László Mátyás
Patrick Sevestre
Editors

ADVANCED STUDIES IN THEORETICAL AND APPLIED ECONOMETRICS 46

The Econometrics of Panel Data

Fundamentals and Recent Developments in Theory and Practice

Third Edition

Springer

Advanced Studies in Theoretical and Applied Econometrics

Volume 46

The titles published in this series are listed at the end of this volume.

László Mátyás • Patrick Sevestre (Eds.)

The Econometrics of Panel Data

Fundamentals and Recent Developments in Theory and Practice

Third Edition

With 13 Figures and 43 Tables

Springer

Editors

Prof. László Mátyás
Central European University
Department of Economics
Nádor u. 9
1051 Budapest
Hungary
matyas@ceu.hu

Prof. Patrick Sevestre
Université Paris 1-Panthéon Sorbonne
Ecole Economique de Paris
(Paris School of Economics)
106-112 Boulevard de l'Hôpital
75013 Paris
France
sevestre@univ-paris1.fr

ISBN: 978-3-540-75889-1 e-ISBN: 978-3-540-75892-1

Advanced Studies in Theoretical and Applied Econometrics ISSN: 1570-5811

Library of Congress Control Number: 2007940369

2nd edition was published by Springer Netherlands in 1995

Cover design: eStudio Calamar S.L., F. Steinen-Broo, Pau/Girona, Spain

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

The aim of this third, completely re-written, re-edited and considerably expanded, edition of this book is to provide a general overview of both the basics and recent, more sophisticated, theoretical developments in panel data econometrics. It also aims at covering a number of fields of applications where these methods are used for improving our knowledge and understanding of economic agents' behaviors. Since the pioneering works of Edwin Kuh (1959), Yair Mundlak (1961), Irving Hoch (1962), and Pietro Balestra and Marc Nerlove (1966), the pooling of cross sections and time series data has become an increasingly popular way of quantifying economic relationships. Each series provides information lacking in the other, so a combination of both leads to more accurate, reliable and informative results than would be achievable by one type of series alone. Over the last three decades of the last century, much fundamental work has been done: investigation of the properties of different estimators and test statistics, analysis of dynamic models and the effects of eventual measurement errors, etc.

The more recent years and in particular the ten years elapsed since the second edition of this book have witnessed even more considerable changes. Indeed, our ability to estimate and test nonlinear models have dramatically improved and issues such as the unobserved heterogeneity in nonlinear models, attrition and selectivity bias have received considerable attention. This explains why the number of chapters dealing with such issues has increased in this third edition. Other recent and important developments relate to the issue of unit roots and cointegration in long times series panels as well as that of cross-sectional dependence that occur in particular in spatial models, and else.

The first objective of this book, which takes up Parts I and II, is to give a complete and state of the art presentation of these theoretical developments. Part I is concerned with the basic fixed effects, random effects and random coefficients models, both linear and nonlinear; Part II deals with various extensions: dynamic models with small T panels, dynamic models with large T panels, models with other sources of endogeneity (measurement errors, simultaneity) and also provides an overview of recent developments in several other directions: attrition and selection bias, pseudo-panels, semi- and non-parametric methods, the Bayesian approach to panel data, the

poolability of individuals, duration models and point processes, and count data models. The second objective of this volume is to provide insights into the use of panel data in empirical studies. Since the beginnings, interest in panel data has mostly been empirically motivated. Panel data methods have gained an increased importance over time and are now applied in a very large spectrum of economic studies. Part III thus deals with studies in several major fields of applied economics, such as foreign direct investments, production frontiers, linked employer-employee data, labor supply, policy analysis and transitions on the labor market. Some of the chapters in this third edition are revised versions of those already published in the previous ones, while several others are completely new contributions. In this respect, we are particularly happy to welcome aboard our new authors. Their input definitely helped to substantially improve the contents of this volume.

The double emphasis of this book (theoretical and applied), together with the fact that all the chapters have been written by well-known specialists in the field, encourage us to hope that it has now become a standard reference textbook for all those who are concerned with the use of panel data in econometrics, whether they are advanced students, professional economists or researchers. The editors have tried to standardize the notation, language, depth, etc. in order to present a coherent book. However, each chapter is capable of standing on its own as a reference in its own topic.

We must address our thanks to all those who have facilitated the creation of this book: the contributors who produced quality work, then took part in an internal refereeing process to ensure a high overall standard; Kluwer Academic Publishers, who had the foresight to publish in a subject which, at the time of the first edition, had a limited, but expanding, audience; and of course Springer which has become our publisher by now. In particular, Cathelijne van Herwaarden, Herma Drees, Marie Sheldon, Martina Bihn, Ruth Milewski and Isabelle George must be thanked for their help in the realization of this volume. Also, the University of Paris–Val de Marne in France; the Monash Research Fund and the Australian Research Council in Australia, and the Budapest University of Economics and the Hungarian Research Fund (OTKA) in Hungary must be thanked for having provided financial support to the editors for the earlier editions. This third edition has benefited from generous financial support provided by the Central European University and the University of Paris-Val de Marne.

Budapest and Paris
January 2008

László Mátyás
Patrick Sevestre

Contents

Part II Advanced Topics

List of Contributors

Jaap H. Abbring
Department of Economics, VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, and Tinbergen Institute,
e-mail: jabbring@econ.vu.nl

John M. Abowd
School of Industrial and Labor Relations, Cornell University, Ithaca, NY 14850, USA,
e-mail: john.abowd@cornell.edu

Chunrong Ai
Department of Economics, University of Florida, Warrington College of Business Administration, 224 MAT, P.O. Box 117140, Gainesville, FL 32611-7140, USA,
e-mail: chunrong.ai@cba.ufl.edu

Luc Anselin
School of Geographical Sciences, Arizona State University, Tempe, AZ 85287, USA,
e-mail: luc.anselin@asu.edu

Pietro Balestra[†]
Faculty of Economics, University of Lugano, via Giuseppe Buffi 13, CH-6904 Lugano, Switzerland.

Badi H. Baltagi
Center for Policy Research, 426 Eggers Hall, Syracuse University, Syracuse, NY 13244-1020, USA,
e-mail: bbaltagi@maxwell.syr.edu

Erik Biørn
Department of Economics, University of Oslo, P.O. Box 1095, Blindern, 0317 Oslo, Norway,
e-mail: erik.biorn@econ.uio.no

Pierre Blanchard
Erudite, Faculté de Sciences Economiques et de Gestion, Université Paris XII Val de Marne, 61 Av. du Général de Gaulle, 94010 Créteil Cédex, France,
e-mail: blanchard@univ-paris12.fr

Rachid Boumahdi
Toulouse School of Economics, GREMAQ and LIHRE; Université des Sciences Sociales de Toulouse, 21 Allée de Brienne, 31000 Toulouse, France,
e-mail: boumahdi@univtlse1.fr

Jörg Breitung
University of Bonn, Institute of Econometrics, Adenauerallee 24-42, 53113 Bonn, Germany, e-mail: breitung@uni-bonn.de

Georges Bresson
ERMES (UMR 7181, CNRS), Université Paris II and TEPP (FR 3126, CNRS), Institute for Labor Studies and Public Policies, 12, place du Panthéon, 75230 Paris Cedex 05, France,
e-mail: bresson01@aol.com

Siddhartha Chib
Olin Business School, Campus Box 1133, Washington University in St. Louis, 1 Brookings Dr, St. Louis, MO 63130,
e-mail: chib@wustl.edu

Christopher Cornwell
Department of Economics, University of Georgia, Athens, GA 30602, USA,
e-mail: cornwl@terry.uga.edu

Bruno Crépon
CREST-INSEE (Paris), CEPR (London) and IZA (Bonn), CREST-INSEE,
15 Boulevard Gabriel Péri, 92245 Malakoff Cedex, France,
e-mail: crepon@ensae.fr

Jean-Pierre Florens
Toulouse School of Economics, Institut Universitaire de France, Université des Sciences Sociales de Toulouse, 21 Allée de Brienne, 31000 Toulouse, France,
e-mail: florens@cict.fr

Denis Fougère
CREST-INSEE (Paris), CEPR (London) and IZA (Bonn), CREST-INSEE,
15 Boulevard Gabriel Péri, 92245 Malakoff Cedex, France,
e-mail: fougere@ensae.fr

Carl Gaigné
INRA, UMR1302, SMART, F-35000 Rennes, France,
e-mail: gaigne@rennes.inra.fr

Mark N. Harris
Department of Econometrics & Business Statistics, Monash University, Victoria 3800, Australia,
e-mail: mark.harris@buseco.monash.edu.au

James J. Heckman
Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, USA; American Bar Foundation and Geary Institute, University College Dublin,
e-mail: jjh@uchicago.edu

Bo Honoré
Department of Economics, Princeton University, Princeton, NJ 08544-1021, USA,
e-mail: honore@Princeton.EDU

Cheng Hsiao
University of Southern California and Nanyang Technological University, Department of Economics, University of Southern California, University Park, Los Angeles, California 90089, USA,
e-mail: chsiao@usc.edu

Hubert Jayet
EQUIPPE, University of Science and Technology of Lille, Faculty of Economics and Social Sciences, 59655 Villeneuve d'Ascq Cedex, France,
e-mail: Hubert.Jayet@univ-lille1.fr

Thierry Kamionka
CNRS and CREST-INSEE (Paris), CREST-INSEE 15, Boulevard Gabriel Péri, 92245 Malakoff Cedex, France,
e-mail: kamionka@ensae.fr

Bertrand Koebel
BETA, Université Louis Pasteur, Strasbourg I and IZA, Bonn, 61 Avenue de la Forêt Noire, F67000 Strasbourg, France,
e-mail: koebel@cournot.u-strasbg.fr

Francis Kramarz
CREST-INSEE (Paris), CREST-INSEE, 15 Boulevard Gabriel Péri, 92245 Malakoff Cedex, France,
e-mail: kramarz@ensae.fr

Jayalakshmi Krishnakumar
Department of Econometrics, University of Geneva, 40 Bd. du Pont d'Ave, CH-1211 Geneva 4, Switzerland,
e-mail: jaya.krishnakumar@metri.unige.ch

François Laisney
BETA, Université Louis Pasteur, Strasbourg I and ZEW, Mannheim, 61 Avenue de la Forêt Noire, F67000 Strasbourg, France,
e-mail: fla@cournot.u-strasbg.fr

Michael Lechner
Swiss Institute for Empirical Economic Research (SEW), University of St. Gallen, Varnbühlstr. 14, CH-9000 St. Gallen, Switzerland,
e-mail: Michael.Lechner@unisg.ch

Julie Le Gallo
CRESE, Université de Franche-Comté, 45D Avenue de l'Observatoire, 25030 Besançon Cedex, France,
e-mail: jlegallo@univ-fcomte.fr

Qi Li
Department of Economics, Texas A&M University, College Station, TX 77843-4228, USA,
e-mail: qi@econmail.tamu.edu

Roman Liesenfeld
Department of Economics, Christian-Albrechts-Universität Kiel, Olhshausenstr. 40-60, 24118 Kiel, Germany,
e-mail: liesenfeld@stat-econ.uni-kiel.de

Stéfan Lollivier
INSEE, 18 boulevard Adolphe Pinard, F-75014 Paris, France,
e-mail: stefan.lollivier@insee.fr

Thierry Magnac
Université de Toulouse 1, Toulouse School of Economics, Manufacture des Tabacs, 21, Allée de Brienne, 31000 Toulouse, France,
e-mail: magnac@cict.fr

Jacques Mairesse
CREST-INSEE (Paris), UNU-MERIT (University of Maastricht), and NBER. CREST-INSEE, 15 boulevard Gabriel PERI, 92245 MALAKOFF cedex, France,
e-mail: mairesse@ensae.fr

Claude Mathieu
Erudite, Faculté de Sciences Economiques et de Gestion, Université Paris XII Val de Marne, 61 Av. du Général de Gaulle, 94010 Créteil Cédex, France,
e-mail: mathieu@univ-paris12.fr

László Mátyás
Central European University, Department of Economics, Nádor u. 9, 1051 Budapest, Hungary,
e-mail: matyas@ceu.hu

Michel Mouchart
Emeritus Professor of Statistics and Econometrics, Institut de statistique, 20 Voie du Roman Pays, B-1348 Louvain-La-Neuve (Belgium),
e-mail: Michel.Mouchart@uclouvain.be

Marc Nerlove
Department of Agricultural and Resource Economics, University of Maryland, 2200 Symons Hall, College Park, MD 20742-5535, USA,
e-mail: mnerlove@arec.umd.edu

Hashem Pesaran
Cambridge University and USC, Sidgwick Avenue, Cambridge, CB3 9DD, United Kingdom,
e-mail: mhp1@econ.cam.ac.uk

Alain Pirotte
ERMES (UMR 7181, CNRS), Université Paris II and TEPP (FR 3126, CNRS), Institute for Labor Studies and Public Policies, 12, place du Panthéon, 75230 Paris Cedex 05, France,
e-mail: apirotte@aol.com

Winfried Pohlmeier
University of Konstanz, Department of Economics, Box D124, D78457 Konstanz, Germany,
e-mail: winfried.pohlmeier@uni-konstanz.de

Jean-François Richard
Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260, USA,
e-mail: fantin@pitt.edu

Peter Schmidt
Department of Economics, Michigan State University, East Lansing, MI 48824, USA,
e-mail: schmidtp@msu.edu

Patrick Sevestre
Université Paris 1 – Panthéon Sorbonne, Ecole Economique de Paris (Paris School of Economics), 106-112 Boulevard de l'Hôpital, 75013 Paris, France,
e-mail: sevestre@univ-paris1.fr

Matthias Staat
University of Mannheim, Department of Economics, D68131 Mannheim, Germany,
e-mail: matthias@pool.uni-mannheim.de

Alban Thomas
Toulouse School of Economics, INRA; Université des Sciences Sociales de Toulouse, 21 Allée de Brienne, 31000 Toulouse, France,
e-mail: thomas@toulouse.inra.fr

Francis Vella
Department of Economics, Georgetown University, Washington DC, USA,
e-mail: fgv@georgetown.edu

Marno Verbeek
Department of Financial Management, RSM Erasmus University, Burg. Oudlaan 50, 3062 PA Rotterdam, The Netherlands,
e-mail: mverbeek@rsm.nl

Frank Windmeijer
Department of Economics, University of Bristol, 8 Woodland Road, Bristol BS8 1TN, UK,
e-mail: f.windmeijer@bristol.ac.uk

Simon Woodcock
Department of Economics, Simon Fraser University, Burnaby, BC, Canada,
e-mail: simon_woodcock@sfu.ca

Part I
Fundamentals

Chapter 1
Introduction

Marc Nerlove, Patrick Sevestre and Pietro Balestra

1.1 Introduction

In his famous and influential monograph, *The Probability Approach in Econometrics*, Haavelmo (1944) laid the foundations for the formulation of *stochastic* econometric models and an approach which has dominated our discipline to this day. He wrote:

> … we shall find that two individuals, or the same individual in two different time periods, may be confronted with exactly the same set of specified influencing factors [and, hence, they have the same y^*, …], and still the two individuals may have different quantities y, neither of which may be equal to y^*. We may try to remove such discrepancies by introducing more "explaining" factors, x. But, usually, we shall soon exhaust the number of factors which could be considered as *common* to all individuals, and which, at the same time, were not merely of negligible influence upon y. The discrepancies $y - y^*$ for each individual may depend upon a great variety of factors, these factors may be different from one individual to another, and they may vary with time for each individual (Haavelmo, 1944, p. 50).

And further that:

> …we find justification for applying them [stochastic approximations] to economic phenomena also in the fact we usually deal only with—and are interested only in—total or average effects of many individual decisions, which are partly guided by common factors, partly by individual specific factors… (Haavelmo, 1944, pp. 51 and 56)

Marc Nerlove
Department of Agricultural and Resource Economics, University of Maryland, 2200 Symons Hall, College Park, MD 20742-5535, USA, e-mail: mnerlove@arec.umd.edu

Patrick Sevestre
Université Paris 1 – Panthéon Sorbonne, Ecole Economique de Paris (Paris School of Economics), 106-112 Boulevard de l'Hôpital, 75013 Paris, France, e-mail: sevestre@univ-paris1.fr

Pietro Balestra
Faculty of Economics, University of Lugano, via Giuseppe Buffi 13, CH-6904 Lugano, Switzerland

This introductory chapter is dedicated to the memory of our late colleague, Pietro Balestra. This chapter largely draws on his and Nerlove's earlier introduction to the second edition of this volume.

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

Marschak (1950, 1953) further amplified Haavelmo's themes in his introduction to Cowles Commission Monographs 10 and 14, observing that:

> The numerous causes that determine the error incurred … are not listed separately; instead their joint effect is represented by the probability distribution of the error, a random variable (1950, p. 18) [which] … is called 'disturbance' or 'shock,' and can be regarded as the joint effect of numerous separately insignificant variables that we are unable or unwilling to specify but presume to be independent of observable exogenous variables. (1953, p. 12).

In this introduction we examine how the basic principle underlying the formulation of econometric models has been carried forward in the development of econometric models and methods for the analysis of panel data. We argue that while fixed effects models may be appropriate in cases in which a population is sampled exhaustively (e.g., data from geographic regions over time) or in which it is desired to predict individual behavior (e.g., the probability that a given individual in a sample will default on a loan), random effects models are more consistent with Haavelmo's view, quoted above, that the "population" we model in econometrics consists not of an infinity of individuals, in general; but of an infinity of decisions. This is not to say, however, that fixed effects models may not be extremely useful as an analytic device.

Moreover, we shall argue, taking a leaf from Knight (1921), that what differentiates the individuals, who make the decisions with which we are concerned, is largely historical, the "three great accumulating funds of inheritance from the past, material goods and appliances, knowledge and skill, and morale." This view has important implications for the relevance and appropriateness of many of the models and methods for the analysis of panel data which have been developed over the past 40 years. We briefly review these developments here and conclude that not only are random effects models most relevant and appropriate but that often our central analytical and modelling concerns are also dynamic. Thus, the most fruitful developments in this enormous literature have been those which deal with the central issues of history and dynamics.

1.2 Data, Data-Generating Processes (DGP), and Inference

In most applications of statistical analysis in the so-called "hard" sciences, the process by which the observed data are generated is transparent, having usually been determined by the investigator by design. In contrast, in many applications in the social sciences, especially in economics, the mechanism by which the data are generated is opaque. In such circumstances, estimation of the parameters of the statistical or econometric model and the testing of specific hypotheses about it are only half the problem of inference. Understanding the process by which the observations at hand are generated is of equal importance. Were the data for example obtained from a sample of firms selected by stratified random sampling from a census of all firms in the United States in 2000? For example, were they obtained from regulatory activity? In the case of time series, the data are almost always "fabricated" in one way or another, by aggregation, interpolation, or extrapolation, or by all three. The nature of the sampling frame or the way in which the data are fabricated must be

part of the model specification on which parametric inference or hypothesis testing is based. Nonparametric inference imposes fewer restrictions on the specification of the DGP but incorporation of knowledge about the nature of the DGP is of equal importance. Almost all the methodological papers in Parts 1 and 2 of this volume focus primarily on problems of estimation and inference from a parametrically well-specified model of how the observed data were generated and variously draw their interest from different types of data and their DGPs. In this section we address the issue of why the DGP matters in the context of a specific, although somewhat abstract, example.

Suppose a longitudinal household survey in which the same households are questioned over time about their actions in, say, a number of consecutive months or years and, initially, about various demographic and economic characteristics. These households differ in various ways, some of which we observe and many which we do not. Some of these differences are the result of their past behavior or past circumstances (path dependence), some are differences in tastes or other unobserved characteristics which may be assumed to be permanent (individual heterogeneity), and some are due to peculiarities not permanently associated with time or individual.[1]

What, in the context of these data, can be considered as random, what is the population from which we may consider the data a sample, and what is a parameter, and what a random variable? These issues are central to an understanding of the DGP.

Statistical and, a fortiori, econometric analysis, are usually based on the idea of sampling from a population in order to draw inferences for the underlying population. But what is the population from which economic data may be supposed to be a sample? In his famous 1944 monograph, Haavelmo (1944, p. 56) wrote, "... the class of populations we are dealing with does not consist of an infinity of different individuals, it consists of an infinity of possible decisions which might be taken ...". In their recent text, *Econometric Theory and Methods*, Davidson and MacKinnon (2004, pp. 30–31) make the same point: "In econometrics, the use of the term population is simply a metaphor. A better concept is that of a data-generating process, or DGP. By this term, we mean whatever mechanism is at work in the real world of economic activity giving rise to the numbers in our samples, that is, precisely the mechanism that our econometric model is supposed to describe. A DGP is thus the analog of a population in biostatistics. Samples may be drawn from a DGP just as they may be drawn from a population. In both cases, the samples are assumed to be representative of the DGP or population from which they are drawn."

What is a random variable in this context and what is not? Whether or not a particular variable can be considered a random draw from some population or not, in principle can be decided by applying the principle of "exchangeability" introduced by de Finetti (1930). In a nutshell, the idea, very Bayesian in flavor, is to ask whether we can exchange two elements in a sample and still maintain the same subjective distribution. Thus, in a panel study of households, are any two households in the sample exchangeable without affecting the distribution, from which we imagine

[1] In his paper, "Identifying the Hand of the Past: Distinguishing State Dependence from Heterogeneity", Heckman (1991) argues that in general it is not possible to distinguish. The ability to do so rests critically "on maintaining explicit assumptions about the way in which observables and unobservables interact."

household observables and unobservables to be drawn? In a panel of state data, are California and Maryland exchangeable without affecting the subjective distribution of the state effects? It's a dicey question – sometimes.

From the standpoint of a Bayesian there is no real distinction between a parameter and a random variable, but in this context we could say that a parameter is an unobserved variable which affects the distribution of the random variables of the model and is unaffected by the particular values such variables take on. It is what we wish to estimate and about which we wish to make inferences. A related concept is that of an exogenous variable. But note here that such an exogenous variable is still a random variable and not a parameter.

In general, in the formulation of econometric models (i.e., the DGP for the process yielding the particular set of data we want to "explain") the distinction between what can be observed and what is not is fundamental. Linear functions are often used to describe such a DGP. To get more precisely to the issues posed by the formulation of the DGP for a sample of economic data, we need to include several observable variables. Suppose that we draw a random sample of N individuals over T time periods; for example a household survey in which we collect observations on the income, x_{it} and consumption of household i, y_{it}, for many households N, in year t over a brief period T. From the survey we have observations on the pairs (x_{it}, y_{it}). Since the households are chosen at random for the survey, but the years over which they are observed are not, the lists $(x_{i1}, y_{i1}, \ldots, x_{iT}, y_{iT})$, $i = 1, \ldots, N$, are exchangeable, but the order within each list is not.

Imagine we are estimating a consumption function and assume a linear relationship subject to error:

$$y_{it} = a + b\, x_{it} + \varepsilon_{it} \tag{1.1}$$

This would be the case, if for example, the joint distribution of variables could be assumed normal and we were trying to estimate the mean of y_{it} for a particular year t conditional on x_{it}. We might then write ε_{it} as

$$\varepsilon_{it} = \mu_i + \lambda_t + u_{it} \tag{1.2}$$

where ε_{it} is an unobserved random variable which is the sum of three effects, all of which are also unobserved: λ_t is a year effect, arguably nonrandom and therefore a parameter to be estimated for each year, t; μ_i is a household effect, which, in view of the way the observations are drawn, should surely be treated as random, and, finally, u_{it} is a random variable to represent all the rest.

We are far from done yet, however. The question remains as to what we should assume about the observable variables, x_{it}. They are clearly random variables jointly distributed with the variable y_{it}. If not subject to errors of measurement, an assumption difficult to justify in the context of an economic survey, are they also independent of, or at least uncorrelated with, the disturbances ε_{it} in (1.1)? This question clearly affects not only what we can say about the DGP which generates our observations, but also how many and what parameters must be considered. Let us examine the regression with some care. Since λ_t is not a random variable but a parameter, consider it to be a constant for each t and add it to the constant a in the regression equation (1.1):

$$y_{it} = a_t^* + bx_{it} + v_{it} \tag{1.3}$$

where

$$\begin{aligned} a_t^* &= a + \lambda_t \quad \text{and} \\ v_{it} &= \mu_i + u_{it} \ . \end{aligned}$$

Suppose that $(v_{i1},\ v_{i2}, \ldots,\ v_{iT})$ is distributed with mean zero and variance–covariance matrix Σ. If x_{it} is strictly exogenous in the regression (1.3), which means

$$E(v_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}) = 0, \ \ \forall i \text{ and } t \ , \tag{1.4}$$

then (1.3) is the usual panel model. This means that b can be estimated by GLS or ML with a dummy variable for each t. Weak exogeneity is a related concept, introduced by Engle, Hendry and Richard (1983). In the context of the regression (1.3), we say x_{it} is weakly exogenous if v_{it} is distributed independently of $\{x_{is}, y_{is}$, for all i and $s \leq t-1\}$, if the marginal distribution of $\{x_{is}, y_{is}$, for all i and $s \leq t-1\}$ does not depend on any unknown parameters in Σ or on b or the λ's, nor does the pdf of $x_{it} \mid \{x_{is}, y_{is}$, for all i and $s \leq t-1\}$. If regression (1.3) satisfies the conditions of strict exogeneity, the likelihood function for the whole sample of observations on x and y factors into two pieces, one of which is the usual regression likelihood and the other is a function of x but not of the parameters in Σ or b or the λ's. In that sense we can treat the observations on x as fixed.

But is exogeneity, weak or strict, a reasonable assumption? Here is what Wooldridge (2002, p. 252) says: "Traditional unobserved components panel models take the x_{it} as fixed. We will never assume the x_{it} are nonrandom because potential feedback from y_{it} to x_{is} for $s > t$ needs to be addressed explicitly."

The assumption that the explanatory variables in the regression are exogenous is generally impossible. If the vector of explanatory variables includes any lagged values of y_{it}, either explicitly or implicitly, the strict or weak exogeneity is generally impossible. Any meaningful DGP describing individual economic behavior is intrinsically dynamic in the sense that the "hand of the past," whether as a result of path dependence or of individual heterogeneity, is ever present. To put the point more explicitly, if, among the observed variables are any initial conditions related to past values of the observed y_{it}'s or to unobservables affecting present and past behavior, at least one of the components of x_{it} must be correlated with ε_{it}. A Hausman test will reject exogeneity of the x's almost certainly. A rejection of exogeneity does not, of course, imply that the unobserved components ε_{it} of the errors in (1.3) are not random (RE) but fixed (FE).

Unfortunately, as Hsiao (2007) points out, this leaves the econometrician between Scylla and Charybdis: We're damned if we do assume that the errors are random, and damned if we don't. Although the RE model avoids the classic incidental parameters problem (Neyman and Scott, 1948), it is necessary in order to obtain unbiased estimates to specify the conditional distribution of μ_i given the x_{it}, and the μ_i are unobservable. As Heckman, quoted above, says, one must be willing

to make "explicit assumptions about the way in which observables and unobservables interact." But most econometricians are not willing to specify such interactions as part of the DGP. Hence, the random effects are treated as parameters rather than random variables. They are viewed as incidental parameters and the object is to get rid of them without distorting the estimates of the structural parameters. There is no universally accepted way of doing so in all contexts, especially not in explicitly dynamic or nonlinear contexts, and, in our view no right way of doing so.

1.3 History and Dynamics

The fundamental fact about society as a going concern is that it is made up of individuals who are born and die and give place to others; and the fundamental fact about modern civilization is that, as previously quoted from Knight (1921), it is dependent upon the utilization of three great accumulating funds of inheritance from the past, material goods and appliances, knowledge and skill, and morale. Besides the torch of life itself, the material wealth of the world, a technological system of vast and increasing intricacy and the habituations which fit men for social life must in some manner be carried forward to new individuals born devoid of all these things as older individuals pass out.

The moral of Knight's characterization is that history is important and individuals have histories. We illustrate our general view of the central principle involved using a simple illustrative example drawn from a paper of Mátyás and Rahman (1992).

Let i index individuals and t time periods. Suppose the relationship we are interested in estimating is

$$y_{it} = \sum_{s=0}^{\infty} \beta_s x_{i,t-s} + \varepsilon_{it} \ . \tag{1.5}$$

The variable x_{it} is assumed to be exogenous and distributed independently of the true disturbances ε_{it} for all finite subsets of the t-index set. We also assume, despite our previous injunction, that

$$\begin{aligned} E(\varepsilon_{it}) &= 0, \ \ \forall i,t \\ E(\varepsilon_{it}\varepsilon_{i't'}) &= \sigma_\varepsilon^2 \quad \text{for} \quad i = i' \quad \text{and} \quad t = t' \\ &= 0 \text{ otherwise} \ . \end{aligned} \tag{1.6}$$

To guarantee some stability in the relationship we are interested in estimating, we must also assume some convergence properties for the sequence of distributed lag weights. Although stronger than necessary, assume they are square-summable:

$$\sum_{s=0}^{\infty} \beta_s^2 < \infty \ . \tag{1.7}$$

Of course, as Mátyás and Rahman note, (1.5) is not estimable with a finite amount of data. Indeed, the time dimension is likely to be very short. Instead, we truncate:

$$y_{it} = \sum_{s=0}^{k} \beta_s x_{i,t-s} + \sum_{s=k+1}^{\infty} \beta_s x_{i,t-s} + \varepsilon_{it} \equiv \sum_{s=0}^{k} \beta_s x_{i,t-s} + \mu_i + \varepsilon_{it} \ . \tag{1.8}$$

Equation (1.8) is in the form of a frequently used random effect model, except that now the individual-specific effects are interpreted in terms of the past histories of each individual in the panel prior to the time when observation begins. Moreover, the assumption that x_{it} is stochastic, although exogenous, is not innocuous. The implications are:

First, interpreting μ_i as fixed, nonstochastic, is not appropriate. If you accept Haavelmo's view that the class of populations which we imagine (1.8) reflects, consists of decisions rather than identifiable specific individuals, then, in principle, we should not even condition on μ_i. However, an exception to this rule is if, for the particular sample of individuals we have drawn (now we can specifically identify each), we want to *predict* future values of y_{it} *for that individual*.

Second, since the x_{it} are themselves considered to be stochastic, for each individual their values over time will in general be correlated. There may also be correlations among x_{it}'s for different values of i if different individuals have some characteristics in common. But we neglect this possibility here. It follows that μ_i and the values x_{it} observed are correlated. Suppose, for example,

$$x_{it} = \rho_i x_{i,t-1} + \omega_{it} \ , \tag{1.9}$$

where $|\rho_i| < 1$ and $E(\omega_{it}) = 0, E(\omega_{it}\omega_{i't'}) = \sigma_i^2$, $i = i'$ and $t = t'$, and $E(\omega_{it}\omega_{i't'}) = 0$, $i \neq i'$ or $t \neq t'$, for all i and t. Let $S = \{0, 1, \ldots, K\}$ be the set of indices for which, given i, x_{it} is observed (normally k will be chosen much less than K). Since

$$E(x_{it}) = 0 \tag{1.10}$$

$$\text{and } E(x_{it}x_{i,t-\tau}) = \frac{\rho_i^{\tau}}{1-\rho_i^2}\sigma_i^2 \ , \tag{1.11}$$

it follows that x_{it}, $t \in S$, will be correlated with μ_i, and the correlation will depend on how close to the beginning of the sample period the observation on x_{it} is taken:

$$E(x_{i\tau}\mu_i) = \sum_{s=k+1}^{\infty} \beta_s E(x_{i\tau}x_{i,t-s}) = \frac{\sigma_i^2}{1-\rho_i^2} \sum_{s=k+1}^{\infty} \beta_s \rho_i^{|\tau-s|} \ , \tag{1.12}$$

for $\tau \in S$. Clearly, this makes the likelihood of the sample much more difficult to determine and introduces some of the parameters, namely β_s, into the relationship between the individual-specific disturbances in (1.8) and the observed past values of the explanatory exogenous variable (we would perhaps be willing to regard σ_i^2 and ρ_i as nuisance parameters).

The important point about this admittedly unrealistic example is that it shows that an entirely new set of questions must be considered. In particular, the error which we make by treating μ_i as independent of the observed values of x_{it} now depends in a complex way on the way in which the distributed lag parameters of interest interact with the nuisance parameters σ_i^2 and ρ_i. Indeed, matters become even more

interesting when we note that the unconditional variance of x_{it} is $\sigma_i^2/(1-\rho_i^2)$, so that, in general the greater σ_i^2 the greater is the signal to noise ratio in (1.8), on the one hand, but, ceteris paribus, the greater is the dependence between $x_{i\tau}$ and μ_i, especially for τ near the beginning of the observation period. Other questions we must ask ourselves are: How can we optimally rid ourselves of the nuisance parameters? How badly does a method, which is based on the assumption that μ_i and the observed $x_{i\tau}$ are uncorrelated, approximate the true ML estimates? What constitute appropriate instruments in considering alternative methods to ML? And so forth.

Consider now an autoregressive model:

$$y_{it} = \alpha y_{i,t-1} + x'_{it}\beta + \mu_i + \varepsilon_{it}, \;\; i = 1, \ldots, N, \;\; t = 1, \ldots, T \; . \tag{1.13}$$

If only semi-asymptotics on N is considered (T finite), we need not assume $\mid \alpha \mid < 1$. On the other hand, the process generating the initial observations is very important. As suggested above, this means that the individuals' past history with respect to both the observed variables x and the latent variables ε becomes crucial.

We can rewrite (1.13) as

$$y_{it} = \alpha^t y_{i0} + \sum_{j=0}^{t} \alpha^j x'_{i,t-j}\beta + \frac{1-\alpha^t}{1-\alpha}\mu_i + v_{it} \tag{1.14}$$

where

$$v_{it} = \sum_{j=0}^{t-1} \alpha^j \varepsilon_{i,t-j} \; .$$

Thus, each observation on the dependent variable y_{it} can be written as the sum of four terms:

The first, $\alpha^t y_{i0}$, depends on the initial values which, as long as T is finite, do influence the behavior of any estimators. Moreover, there is no good reason (as in Balestra and Nerlove, 1966) to assume that these are fixed (to condition upon their values) and independent of individual specific effects. Indeed, unless there is something special about the initial date of observation, there is no justification for treating the initial observation differently from subsequent observations or from the past, but unobserved, previous values.

The second term in (1.14) depends on the current and past values of the exogenous variables x'_{it}. The form that this dependence takes depends not only on the dynamics of the model, but also on the way in which individuals' past histories differ (Knight, 1921).

The third term depends on remaining individual specific effects which are assumed to be orthogonal to the individual's past history.

Finally, the last term is a moving average in past values of the remaining disturbances, which may also be written:

$$\begin{aligned} v_{it} &= \alpha v_{it} + \varepsilon_{it} \quad t \geq 1 \\ v_{it} &= 0 \quad t = 0 \; . \end{aligned}$$

Conditioning on the initial observations implies that they can be treated as fixed constants independently of μ_i and ν_{it}. They need not be independent of any of the lagged values of the explanatory x's *which are included.* But if any truncation within-sample occurs, the truncation remainder will be part of the individual specific disturbance, as shown above, and thus the initial values of the endogenous variable are not independent of the disturbance and cannot be treated as fixed.

This point can be made in another way (following Chap. 8): Write the cross section of initial observations as a function of past x's, μ_i, and ε_{i0},

$$y_{i0} = f(x'_{i,0}, x'_{i,-1}, \ldots, \mu_i, \varepsilon_{i0}) \,. \tag{1.15}$$

The problem is now related to whether or not we choose to regard μ_i as fixed or random. *If μ_i is fixed* and thus independent, cross-sectionally, of ε_{i0}, and if $x'_{i,t-j}$ $j = 0, 1, \ldots$, are cross-sectionally exogenous, then the y_{i0} can be conditioned on. They are still, however, random variables. *But, if the μ_i are random variables*, the y_{i0} are not exogenous. This shows that in a dynamic context fixed effects versus error components assumptions make a big difference. Our preceding argument suggests that the error components assumption is the more appropriate.

In this case, the literature suggests a variety of different assumptions about the initial observation leading to different optimal estimation procedures and implying different properties for suboptimal estimates. One line takes the generating process of the initial observations to be different from that of subsequent observations. Anderson and Hsiao (1982), for example, suggest a general form

$$y_{i0} = k_0 + k_1\mu_1 + k_2\varepsilon_{i0} \,. \tag{1.16}$$

If $k_1 = k_2 = 0$, the initial observations are fixed and identical. If $k_0 = k_1 = 0$ and $k_2 \neq 0$, the y_{i0} are random variables independent of the disturbances in (1.13). If $k_0 = 0, k_1 = 1/(1-\alpha)$ and $k_2 = 1/(1-\alpha^2)^{1/2}$ the individual autoregressive processes which generate the y's are stationary, and so forth.

But, although convenient, it is not very reasonable to suppose the initial observation to be generated by a mechanism much different than that which generates subsequent observations. Bhargava and Sargan (1983) suggest

$$y_{i0} = k_0 + x^{*'}_{i0}\gamma + k_1\mu_i + k_2\varepsilon_{i0} \,, \tag{1.17}$$

where the $x^{*'}_{i0}$ are exogenous variables, possibly different from x'_{i0} but quite possibly correlated with subsequent observed x'_{it}'s and where γ may or may not equal β. This formulation obviously encompasses the stricter assumption that the *same* mechanism generates y_{i0} and subsequent y_{it}'s and allows the exogenous variables themselves to be generated by other independent dynamical systems.

Assuming fixed effects in a dynamic framework and estimating them as if they were constants (or eliminating them by taking deviations from individual means) together with the autoregressive coefficient α leads to inconsistent estimates of the latter. This was noted in Nerlove (1971), Nickell (1981) and proved by Sevestre and Trognon (1985). Although $y_{i,t-1}$ and ε_{it} are uncorrelated, their

respective individual means are correlated with each other, with ε_{it} and with $y_{i,t-1}$. Instrumental variable methods have been proposed to get around this problem (e.g., Balestra and Nerlove, 1966), but as shown in Nerlove (1971), they can result in very erratic estimates if the instruments themselves have relatively low explanatory value.

Conditioning on the initial values of the endogenous variable also leads to troublesome problems. As noted in Nerlove (1971), the estimates of α appear to be inconsistent even when an error components model is assumed and σ_μ^2 and σ_ε^2 are estimated together with other parameters of the model. This was proved in Trognon (1978). Bhargava and Sargan (1983) show that this does not happen when the likelihood function is unconditional, i.e., when it takes into account the density function of the first observation, e.g. as determined by (1.17) and assumptions about the k's, γ, and the densities of μ_i and ε_{i0}. Our opinion on this matter is that it is most plausible and appropriate to assume that the mechanism which generates the initial observation is highly similar, if not identical, to that which generates subsequent observations. If observations on past values of the exogenous variables are not generally available, it would be preferable to model their independent determination rather than to assume their joint effect, $x_{i0}^{*'}\gamma$, to be fixed constants. At least, such an approach would be more consistent with Haavelmo's views as quoted above.

When the solution to the likelihood equations is not on a boundary and when the likelihood function is locally concave at such a solution, the solution with the largest value is consistent, asymptotically efficient, and root-N asymptotically normally distributed with variance-covariance matrix equal the inverse information matrix. Provided the marginal distribution of the initial values y_{i0}, can be correctly specified, the unconditional density of $y_{iT}, \ldots, y_{i0}$, conditional only on the values of observed exogenous variables, gives rise to a likelihood function which has an interior maximum with probability one. If the marginal density of the initial values is misspecified, ML estimates are no longer consistent.

It is not, in fact, difficult to obtain the unconditional likelihood function once the marginal distribution of the initial values is specified. The problem is a correct specification of this distribution. Suppose that the dynamic relationship to be estimated is stationary so that $|\gamma| < 1$. Consider (1.14) for y_{i0} and the infinite past:

$$y_{i0} = \sum_{j=1}^{\infty} \gamma^j \beta x_{i-j} + \frac{1}{1-\gamma}\mu_i + v_{i0} \quad \text{where} \quad v_{it} = \gamma v_{it-1} + \varepsilon_{it} \,. \tag{1.18}$$

(Recall that all variables are expressed as deviations from their overall means). If $\beta = 0$, so that the relationship to be estimated is a pure autoregression, the vector of initial values $y_0 = (y_{10}, \ldots, y_{N0})$ has a joint normal distribution with means 0 and variance–covariance matrix $(\frac{\sigma_\mu^2}{(1-\gamma)^2} + \frac{\sigma_\varepsilon^2}{(1-\gamma^2)})I_N$. The unconditional likelihood is therefore

$$\begin{aligned}\log L(\gamma, \sigma_\mu^2, \sigma_\varepsilon^2 | y_{11}, \ldots, y_{NT}; y_{10}, \ldots, y_{N0}) \\ = -\frac{NT}{2}\log(2\pi) - \frac{NT}{2}\log(\sigma^2) - \frac{N}{2}\log(\xi) - \frac{N(T-1)}{2}\log(\eta) \\ -\frac{1}{2\sigma^2}\sum_i\sum_t \left(y_{it}^* - \gamma y_{it-1}^*\right)^2 - \frac{N}{2}\log\left(\frac{\sigma_\mu^2}{(1-\gamma)^2} + \frac{\sigma_\varepsilon^2}{1-\gamma^2}\right) \\ -\left[\frac{1}{2}\left(\frac{\sigma_\mu^2}{(1-\gamma)^2} + \frac{\sigma_\varepsilon^2}{1-\gamma^2}\right)\right]\sum_i y_{i0}^2. \end{aligned} \tag{1.19}$$

where y_{it}^* is obtained by applying the standard GLS transformation to y_{it}. To maximize, express σ_μ^2, σ_ε^2, ξ and η in terms of ρ. For given ρ in the interval $[0,1)$, concentrate the likelihood function with respect to σ^2 and γ. This is a little more complicated than the usual minimization of the sum of squares in the penultimate term because γ enters the final term as well. Then do a gradient search on ρ.

When $\beta \neq 0$, things are more complicated still. Various alternative specifications considered in the literature are reported and analyzed in Chap. 8.[2] Considerable simplification, however, can be obtained if, following Nerlove (1971), we are willing to assume that x_{it} follows a well-specified common stationary time series model for all individuals i.

With these general principles in mind, we now turn to a review of other methodological developments that are considered in this volume. Indeed, since the early work of Mundlak (1961) and Balestra and Nerlove (1966), panel or longitudinal data have become increasingly important in econometrics, and methods for the analysis of such data have generated a vast literature much of which has been summarized in the first two editions of this volume. In the last ten years there has been an extraordinary further growth, captured here in eleven completely new chapters and fifteen significantly revised chapters which appeared in the earlier editions.

1.4 A Brief Review of Other Methodological Developments

The most common model for the analysis of panel data is the linear model in which explanatory variables are taken to be exogenous, that is independent of the disturbances in the equation or, in the case of the random coefficients model, of the distributions of the coefficients. When the coefficients (except for the constant term) in the linear relationship with which we describe the data are assumed to be constant, it is usual to distinguish between fixed effects and error components models.

[2] One interesting possibility discussed there is to choose y_{i0} a linear function of some observed individual specific time invariant exogenous variables and a disturbance which is decomposed as the sum of the individual specific disturbances μ_i and a remainder. The first-order equations for maximizing the likelihood then take on a simple recursive form when $\beta = 0$, and permit other simplification when $\beta \neq 0$. But if we knew some individual specific time invariant observed variable influenced behavior why not incorporate them directly in the equation to be estimated?

In the case of the former, the intercepts are assumed to vary across individuals at the same point in time and, possibly, over time for all individuals taken together. In the case of the latter, the variations are assumed to be random and uncorrelated both with the observed explanatory variables and the latent disturbance in the equation.

A considerable quantity of interesting mathematics has been developed for both types of models. A number of different projection matrices exist, which take deviations between the raw observations and various means, across individuals, across time periods, over all, and of various means from other means. These projections can be used to define different possible estimators in fixed effects models or the spectral decomposition of the disturbance variance–covariance matrix in the case of error components models. A principal result is then the demonstration, first noted by Maddala (1971), that the Generalized Least Squares (GLS) estimators of the slope parameters in the error components case are a weighted combination of estimators in the fixed effects case (the so-called "between" and "within" distinction among possible estimators). See Chaps. 2 and 3.

An important distinction is made between fully asymptotic theory in which the limiting properties of estimators are analysed when both the number of time periods and the number of individuals goes to infinity and semi-asymptotic theory in which the number of individuals (or the number of time observations) is assumed to increase without bound, that is, asymptotics in only one of two dimensions. Clearly, in the case of random effects models, the moments of the distribution of the effect whose dimension is not increased in the calculation cannot be semi-asymptotically consistently estimated.

As long as the model is not dynamic, that is, does not contain a distributed lag, lagged values of the dependent variable, or the equivalent stock or state variable, the GLS estimators of these coefficients have the usual good small sample and asymptotic properties, The problem, then, is that the elements of the disturbance variance–covariance matrix are unknown. Since consistency of the variance components estimates depends on the asymptotics assumed, the usual justification for a two-stage procedure (feasible GLS or FGLS) based on first-stage consistent estimates of the variances and covariances of the panel model disturbances does not clearly apply. Indeed, in some cases the FGLS may not even be consistent.

Various interesting extensions of both the fixed effects and error components linear models have recently been made and are presented in this volume. They deal with:

(a) random coefficient models (Chap. 6) and spatial models (Chap. 19). These are important and rapidly expanding fields. Indeed, allowing behaviors to vary randomly across individuals can be an attractive way to account for heterogeneity. Also, the existence of spatial dependence should clearly not be ignored when dealing with regional or industry-level data where assuming the "individuals" to behave independently from each other is clearly a strong assumption;

(b) linear models with random regressors and the Chamberlain (1984) approach (Chaps. 4 and 5). As emphasized in the previous section, it is indeed heroic to assume the absence of correlation between the individual effects and the regressors,

in particular. Instrumental variables estimators may then be a useful tool in such a context. The chapter devoted to the Chamberlain approach shows that one important advantage of this approach is to permit a unified treatment of both fixed effects and random effects models in such a context.

(c) data with measurement errors and simultaneous equation models (Chap. 10). Inconsistencies resulting from the simultaneity of individuals'decisions are quite well-known and the treatment of such an issue does not need any long justification. Griliches (1986) persuasively argues the need to understand and model the processes generating errors in economic data in the estimation of economic relations. Griliches and Hausman (1986) provide a pioneering application to panel data. Moreover, problems associated with measurement errors are more important than they might seem at first, because of the increasing importance of so-called "pseudo panel" data (Chap. 11) and the application of measurement error models to the analysis of such data as if they were true panel data. For many types of problems true panel data are not available, but rather several cross sections at different points in time are. For example, surveys of consumer expenditures based on a sample of individual households are made every few years in the UK or the US. Surveys to determine unemployment and labor force participation are made monthly on the basis of a rotating sample. Pseudo panel methods for treating such data are described in Chap. 11. These methods go back to Deaton (1985) who proposed dividing the sample into "cohorts" sharing common demographic, socio-economic, or historical characteristics, then treating the "cohort" averages as observations on "representative" individuals in a panel. Because each "cohort" observation is based on a sample of the true population cohort, the averages, treated as observations, contain sampling errors. Thus, Deaton proposed that the observations be considered as measurements of the "true" values with errors.

What should we make of this approach from the standpoint of the fundamental issues of history and dynamics? It goes without saying that we want to make use of whatever data is available in an appropriate way. The question is what do the cohort averages mean and how should relationships among them be interpreted? Deaton's cohorts and his proposed treatment of cohort averages is similar to the notion of a representative economic agent, introduced by Alfred Marshall in the last century, and in widespread theoretical use today. Kirman (1992) has given a detailed critique of the concept and many of his points apply in the present context. Essentially, relationships among averages, or for representative individuals, are often not interpretable directly in terms of individual behavior since the relationships among the aggregates is often a result of the aggregation. Another way of saying the same thing is that the aggregate relationships are reduced forms from which the underlying structural relations (at the individual level) will not generally be identifiable. This is particularly the case when differences among individuals are historical to a significant degree and when the relationships of interest are dynamic. To the extent that the cohort-defining variables succeed in classifying individuals together who share common histories and exhibit common forms of (dynamic) behavior, the use of pseudo panel data as if they were true panel data subject to sampling error will be successful. But to the extent that unobserved heterogeneity in either respect remains,

the relationships obtained from pseudo panel data may not permit identification of the underlying structure of interest.

(d) dynamic models are considered in Chaps. 8 and 9. As we have argued above, most relationships of interest are likely to be dynamic and the past histories of individuals are almost always important determinants of current behavior. While Chap. 8 considers the case where the number of periods of observation is finite, Chap. 9 considers the situation where this number of periods can be seen as large enough to consider a T-asymptotics; stationarity of the DGP and the existence of cointegation relationships between variables have to be considered. Let us go back for a while to the former context (finite T). The GLS estimates in an error components setting are obtained by transforming the observations to weighted sums of Between and Within variances and covariances, using appropriate weights based on the two distinct characteristic roots of the variance–covariance matrix of the residuals $u_{it} = \mu_i + \varepsilon_{it}$. The covariance matrix of the disturbances can be expressed as

$$\begin{aligned}\sigma^2\Omega &= \sigma^2(\rho(I_N \otimes J_T) + (1-\rho)I_{NT}\}) \\ &= \sigma^2(\rho(I_N \otimes J_T) + (1-\rho)(I\_N \otimes I_T)) \\ &= \sigma^2(I_N \otimes (\rho J_T + (1-\rho)I_T))\,. \end{aligned} \tag{1.20}$$

where $\sigma^2 = \sigma_\mu^2 + \sigma_\varepsilon^2$, $\rho = \sigma_\mu^2/\sigma^2$ and its distinct roots are $\xi = (1-\rho) + T\rho$ and $\eta = (1-\rho)$. Applying this transformation to the dynamic error components specification in (1.13) and replacing $\frac{\eta}{\xi} = \frac{1}{\theta^2} = \lambda$, the normal equations to be solved for the GLS estimates become:

$$\begin{pmatrix} W_{yx} + \lambda B_{yx} \\ W_{yy_{-1}} + \lambda B_{yy_{-1}} \end{pmatrix} = \begin{pmatrix} W_{xx} + \lambda B_{xx} & W_{x,y_{-1}} + \lambda B_{x,y_{-1}} \\ W_{y_{-1},x} + \lambda B_{y_{-1},x} & W_{y_{-1},y_{-1}} + \lambda B_{y_{-1},y_{-1}} \end{pmatrix} \gamma\,. \tag{1.21}$$

In this case, the calculated RSS/NT estimates not σ^2 but $\eta\sigma^2$. As Maddala (1971) points out, the GLS estimates with $\lambda = \frac{1}{\theta^2}$ can be considered members of a more general class of estimators obtained through different choices of λ. Let $\widehat{\gamma}(\lambda)$ be the estimator of γ obtained by solving the above equations for an arbitrary value of λ. Sevestre and Trognon (1985) show that for the case in which $\beta = 0$, the purely autoregressive case, the following inequality holds:

$$\begin{gathered}\widehat{\gamma}(0) \le \gamma \le \widehat{\gamma}(\theta^2) \le \widehat{\gamma}(1) \le \widehat{\gamma}(\infty) \\ \text{i.e., Within} \le \gamma \le \text{GLS} \le \text{OLS} \le \text{Between}\,.\end{gathered} \tag{1.22}$$

Remarkably, therefore, the GLS estimate is inconsistent in this case. The problem is that the lagged dependent variable is correlated even with the transformed disturbance. Since $\widehat{\gamma}(\lambda)$ is a continuous function of λ, there exists a value λ^* in the interval $[0, \theta^2]$ for which $\widehat{\gamma}(\lambda) = \gamma$. In an earlier paper, Sevestre and Trognon (1983) have derived this value. They also show that when $\beta \neq 0$, the estimate $\widehat{\gamma}(\lambda)$ behaves almost the same as in the purely autoregressive case. Since the λ^* estimate is consistent when there are no exogenous variables, it remains so when there are. The trick is to obtain a consistent estimate of λ^* which can be accomplished by finding

an appropriate instrumental variable for y_{-1}. Even in this case the results depend heavily on the distribution of the estimate of λ^*.

In the dynamic error components model, not only are the OLS pooled regression estimates, the fixed effects or Within estimates, and the Between estimates inconsistent, but so are the GLS estimates using the true value of ρ. However, the method of instrumental variables may be used to obtain a feasible member of the λ-class of estimates which is consistent. Unfortunately, this estimate may have a very large variance. The method of choice in most cases is Maximum Likelihood (ML), provided, of course, that associated computational difficulties can be resolved. But even when the matrix of observed regressors is assumed to be nonstochastic, the properties of ML estimators may no longer be fully optimal asymptotically. Although consistent ML estimates of the coefficients of observed exogenous and of the nonspecific residual variance can be obtained either in the asymptotic or the semi-asymptotic sense, consistent ML estimates of the individual specific residual variance cannot be obtained except in the semi-asymptotic sense. In the dynamic case, however, maximum likelihood based on the likelihood function conditional on the initial observation, or more generally the state, can yield inconsistent estimates (Trognon, 1978).

Other developments covered by specific chapters in this third edition consist of the semi-parametric and non-parametric methods that can be used for analyzing panel data (Chap. 14), the Bayesian approach to panel data analysis (Chap. 15), and the question of the poolability of individuals in a panel (Chap. 16).

Chapters 7, 12, 13, 17 and 18 deal with latent variables and other forms of non-linear models in a panel data context. Two points are worth making in this respect: First, it is frequently more difficult to see how elements of individual heterogeneity should be introduced, in contrast to the simple way in which such heterogeneity is introduced in equations (whether linear or not) in terms of *disturbances*. In these non-linear models, even in the case in which all the explanatory variables are truly exogenous, failure to take account of heterogeneity may result in bias, not merely inefficiency, whereas no bias results in the linear case.

The solution in principle is to formulate a model in terms of the probability of individual observations and then to "integrate out" the heterogeneity factors if these can be parametrically specified. In practice, of course, this is rarely possible analytically and may even be extremely difficult computationally. Methods of simulated moments (see McFadden, 1989, and Chap. 13, below) are of considerable utility in this connection.

An important application of latent variables models (which are largely highly nonlinear) is to selection bias and incompleteness in panel data (Chap. 12). In the case of selection bias, a rule other than simple random sampling determines how sampling from the underlying population takes place. Ignoring the nature of the selectivity mechanism may seriously distort the relationship obtained with respect to the true underlying structure. Heckman (e.g., 1990, and references cited therein) has pioneered in this analysis. The greatest problem in panel data in this connection is attrition (sometimes resolved through partial rotation which has its own problems). The probability of nonresponse increases when the same individual is repeatedly sampled. In Chap. 12, it is shown that the crucial question is whether the observed

values in the sample can be considered as the result of a simple random drawing or whether, on the contrary, they are "selected" by some other rule, random or not. In the case of simple random selection, standard estimation and inference are appropriate, and we say the selection rule is ignorable. On the other hand, if selection is nonrandom with respect to factors reflecting heterogeneity, that is correlated with them, standard techniques yield biased estimates and inferences. In this case the selection rule must be explicitly modelled to correct for selection biases. The authors of Chap. 12 show how this can be done for both and for random effects models. Because consistent estimation in the case of a non-ignorable selection rule is much more complicated than in the ignorable case, several tests are proposed to check whether the selection rule is ignorable.

There are other key methodological chapters in this third edition. These include chapters on the use of simulation techniques for estimating panel models (Chap. 13), on the Generalized Method of Moments for count variable models (Chap. 18) and a long chapter on duration models and point processes (Chap. 17).

Finally, Part III of this third edition contains a number of surveys about possible applications of the above methods. Applications of panel data are very diverse, depending, of course, on the availability of such data in specific substantive contexts. This volume contains new chapters on foreign direct investment (Chap. 20), linked employer–employee data (Chap. 22) and policy analysis (Chap. 24). This third part also contains revised versions of previously published chapters about production frontiers and productive efficiency (Chap. 21), labor supply (Chap. 23), labor market transitions (Chap. 25) and a fully updated version of the software review (Chap. 26). In addition to surveying important substantive areas of research these chapters are particularly useful in illustrating our message.

Obviously, panel data (or pseudo panel data) are essential if we want to estimate dynamic relationships at an individual or disaggregated level. As soon as the focus is on dynamics, historically generated heterogeneity becomes a central issue. Models of factor demand (labor and capital investment) reveal the crucial role of expectations. In this connection it is interesting to note the special impact of heterogeneity on expectations. Panel data provide a unique opportunity to study expectation formation and to test various hypotheses about expectation formation (See e.g., Nerlove, 1983, Nerlove and Schuermann, 1995). Often, however, panel data do not contain direct observations on expectations but, as is typically the case with time series data, only on other variables affected by expectations. In this case, we formulate a model of expectation formulation and infer indirectly the parameters of both the behavioral and the expectational model. To see how heterogeneity plays a critical role, it is useful to consider two simple examples: adaptive expectations and rational expectations.

Suppose that the model we wish to estimate is

$$y_{it} = \alpha x^*_{it} + u_{it} \quad i = 1, \ldots, N;\ t = 1, \ldots, T\ , \tag{1.23}$$

where expectations are adaptative:

$$x^*_{it} = \beta x^*_{i,t-1} + (1-\beta) x_{i,t-1} + v_{it}\ . \tag{1.24}$$

Even if the disturbances in the behavioral equation (1.23) are i.i.d. random variables, it is unlikely that past history and past experience will play no latent role in the determination of current expectations, not fully taken into account by $x^*_{i,t-1}$. Thus, write

$$v_{it} = \mu_i + \varepsilon_{it} \ , \tag{1.25}$$

where the individual specific effects are likely to be correlated with past x_{it}'s and also, presumptively, with past u_{it}'s and x^*_{it}'s. The usual transformation of (1.23) and (1.24) then yields

$$y_{it} = \beta y_{i,t-1} + \alpha(1-\beta)x_{i,t-1} + \alpha\mu_i + \alpha\varepsilon_{it} + u_{it} - \beta u_{i,t-1} \ . \tag{1.26}$$

Not only do the usual difficulties, discussed above, arise because of the correlation between $y_{i,t-1}$ and μ_i, but the third term of the disturbance is serially correlated. Moreover, if the individual specific disturbances, μ_i, are correlated with past x_{it}'s, the lagged values of these will no longer serve as instruments.

Still more interesting things happen in the case of rational expectations. In this case (1.24) is replaced by

$$x^*_{it} = E(x_{it} \mid \Omega_{i,t-1}) \ , \tag{1.27}$$

where $\Omega_{i,t-1}$ is the set of information available to the i-th individual at the time when his expectations are formed. In principle, $\Omega_{i,t-1}$ not only contains that individual's own past history, but also observations on aggregates of individuals, and may include knowledge of the way in which individual decisions interact to produce aggregates. For example, supppose

$$z_t = \sum_{i=1}^{N} y_{it} \ . \tag{1.28}$$

Then, for the i-th individual,

$$\Omega_{i,t-1} = \{y_{i,t-1}, \ldots; x_{i,t-1} \ldots; z_{t-1}, \ldots\} \ . \tag{1.29}$$

Rational expectations imply

$$y_{it} = \alpha E(x_{it} \mid \Omega_{i,t-1}) + u_{it} = \alpha E(x_{it} \mid y_{i,t-1}, \ldots; x_{i,t-1}, \ldots) + u_{it} \ . \tag{1.30}$$

Now if the value of x_{it} faced by each individual is a function, peculiar to that individual, of z_t:

$$x_{it} = f_i(z_t) \ , \tag{1.31}$$

which may also be stochastic, then

$$E(x_{it} \mid \Omega_{i,t-1}) = E(f_i(z_t) \mid \Omega_{i,t-1}) \ . \tag{1.32}$$

So, for example, if

$$x_{it} = \gamma z_t + \delta_{it} \ , \tag{1.33}$$

then

$$E(x_{it} \mid \Omega_{i,t-1}) = \gamma E(z_t \mid \Omega_{i,t-1}) + E(\delta_{it} \mid \Omega_{i,t-1}) \,. \tag{1.34}$$

The last term on the right hand side of (1.34) will not generally be zero. Suppose it is. Such a simplification does not essentially affect the nature of the difficulties involved. Then

$$y_{it} = \alpha\gamma E(z_t \mid \Omega_{i,t-1}) + u_{it} = \alpha\gamma E\left(\sum_{i=1}^{N} y_{it} \mid \Omega_{i,t-1}\right) + u_{it} = \alpha\gamma \sum_{i=1}^{N} E(y_{it} \mid \Omega_{i,t-1}) + u_{it} \,. \tag{1.35}$$

Hence, if $u_{it} = \theta_i + \nu_{it}$,

$$E(y_{it} \mid \Omega_{i,t-1}) = \alpha\gamma \sum_{i=1}^{N} \{\alpha\gamma E(y_{it} \mid \Omega_{i,t-1})\} + \sum_{i=1}^{N} E(\theta_i \mid \Omega_{i,t-1}),\ i = 1, \ldots, N \,. \tag{1.36}$$

Equation (1.36) are N equations for each t, which, in principle, can be solved for the N values

$$E(y_{it} \mid \Omega_{i,t-1})$$

in terms of the contents of $\Omega_{i,t-1}$ *for all* N individuals and the sum of expectations

$$\lambda_{t-1} = \sum_{i=1}^{N} E(\theta_i \mid \Omega_{i,t-1}) \,.$$

In general

$$E(y_{it} \mid \Omega_{i,t-1}) = a_1 g_i(\Omega_{i,t-1}) + a_2 \lambda_{t-1} \,.$$

Then we can replace the left hand side of (1.34) by

$$x^*_{it} = a_1 \sum_{i=1}^{N} g_i(\Omega_{i,t-1}) + a_2 N \lambda_{t-1} \,. \tag{1.37}$$

So (1.23) becomes

$$y_{it} = \alpha a_1 \sum_{i=1}^{N} g_i(\Omega_{i,t-1}) + \alpha a_2 N \lambda_{t-1} + u_{it} \,. \tag{1.38}$$

It follows that the appropriate equation now contains a specific time-varying, individual-nonspecific, effect in addition to θ_i and ν_{it}. This effect is correlated with the element in $\Omega_{i,t-1}$ since it is an expectation conditional on $\Omega_{i,t-1}$. Finally, it can be seen that the parameters of g_i, a_1, and a_2 and α are not generally separately identifiable. The bottom line is that, if one believes in rational expectations, one is in deep trouble dealing with panel data.

Unless future values of the exogenous variables are in the information set $\Omega_{i,t-1}$ when expectations are formed, all of the applications discussed in Part III have this problem.

1.5 Conclusion

In this introductory chapter we have tried to bring out the following points:

(a) One of the main reasons for being interested in panel data is the unique possibility of uncovering disaggregate dynamic relationships using such data sets.
(b) In a dynamic context, one of the primary reasons for heterogeneity among individuals is the different history which each has.
(c) If the relevant "population" is, following Haavelmo, the space of possible decisions, different past histories take the form of individual specific random variables which are generally correlated with all of the variables taken as explanatory, not just the lagged values of the endogenous variable. The former therefore cannot be conditioned upon in the usual way.
(d) Finally, although the adaptive expectations model does not introduce any new complications, rational expectations introduce a time specific, individual non-specific, component in the error component formulation, as well as a fundamental failure of identifiability.

Panel data econometrics is one of the most exciting fields of inquiry in econometrics today. Many interesting and important problems remain to be solved, general as well as specific to particular applications. We hope that this volume is the place to start.

References

Anderson, T. W., and C. Hsiao (1982) Formulation and Estimation of Dynamic Models Using Panel Data, Journal of Econometrics, 18, 47–82.

Balestra, P., and M. Nerlove (1966) Pooling Cross-Section and Time-Series Data in the Estimation of a Dynamic Economic Model: The Demand for Natural Gas, Econometrica, 34, 585–612.

Bhargava, A., and D. Sargan (1983) Estimating Dynamic Random-Effects Models from Panel Data Covering Short Time Periods, Econometrica, 51, 1635–1659.

Chamberlain, G. (1984) "Panel Data." Pp. 1247–1318 in Z. Griliches and M. Intriligator, eds. Handbook of Econometrics, Vol. 2. New York: Elsevier.

Davidson, R., and J.G. MacKinnon (2004) Econometric Theory and Methods. New York: Oxford University Press.

de Finetti, B. (1930) Problemi Determinati e Indeterminati nel Calculo delle Probabilità, Rend. R. Acc. Naz. Lincei, Series 6, Vol. 12, fasc. 9.

Deaton, A. (1985) Panel Data from a Time Series of Cross-Sections, Journal of Econometrics, 30, 109–126.

Engle, R.F., D.F. Hendry, and J.-F. Richard (1983). Exogeneity. Econometrica, 51, 277–304.

Griliches, Z. (1986) "Economic Data Issues." Pp. 1465–1514 in Z. Griliches and M. Intriligator, eds. Handbook of Econometrics, Vol. 3. New York: Elsevier.

Griliches, Z., and J.A. Hausman (1986) Errors in Variables in Panel Data, Journal of Econometrics, 31, 93–118.

Haavelmo, T. (1944) The Probability Approach in Econometrics; Supplement to Econometrica, 12.

Heckman, J.J. (1990) Varieties of Selection Bias, American Economic Review, Papers and Proceedings, 80, 313–318.
Heckman, J. (1991) Identifying the Hand of Past: Distinguishing State Dependence from Heterogeneity, The American Economic Review, 81, 75–79.
Hsiao, C. (2007) Panel Data Analysis. Advantages and Challenges, TEST, 16, 1–22.
Kirman, A. (1992) Whom or What Does the Representative Individual Represent? Journal of Economic Perspectives, 6, 117–136.
Knight, F.H. (1921) Risk, Uncertainty and Profit. Boston: Houghton Mifflin Co.
Maddala, G.S. (1971) The Use of Variance Components Models in Pooling Cross-Section and Time Series Data, Econometrica, 39, 341–358.
Marschak, J. (1950) "Statistical Inference in Economics: An Introduction". Pp. 1–50 T.C. Koopmans, ed. in Statistical Inference in Dynamic Economic Model. Koopmans. New York: John Wiley.
Marschak, J. (1953) "Economic Measurements for Policy and Prediction." Pp. 1–26 in W.C. Hood and T.P. Koopmans, Studies in Econometric Method ed.. New York: John Wiley.
Mátyás, L., and S. Rahman (1992) An Alternative Approach for the Estimation of Distributed Lag Models in Panel Data, Working Paper No. 4/92, Department of Econometrics, Monash University.
McFadden, D.L. (1989) A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration, Econometrica, 57, 995–1026.
Mundlak, Y. (1961) Empirical Production Functions Free of Management Bias, Journal of Farm Economics, 43, 44–56.
Nerlove, M. (1971) Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross-Sections, Econometria, 39, 359–382.
Nerlove, M. (1983) Expectations, Plans and Realizations in Theory and Practice, Econometrica, 51, 1251–1280.
Nerlove, M., and T. Schuermann (1995) "Expectations: Are they Rational, Adaptive, or Naive? An Essay in Simulation-Vased Inference." Pp. 354–381 in G.S. Maddala, P. Phillips and T.N. Srinivasan, eds. Advances in Econometrics and Quantitative Economics. Oxford: Basil Blackwell.
Neyman, J., and E. Scott (1948) Consistent Estimates Based on Partially Consistent Observations. Econometrica, 16, 1–32.
Nickell, S. (1981) Biases in Dynamic Models with Fixed Effects, Econometrica, 49, 1417–1426.
Sevestre, P., and A. Trognon (1985) A Note on Autoregressive Error-Components Models, Journal of Econometrics, 29, 231–245.
Trognon, A. (1978) Miscellaneous Asymptotic Properties of OLS and ML Estimators in Dynamic Error Components Models, Annales de l'INSEE, 30–31, 631–658.
Wooldridge, J.M. (2002) Econometric Analysis of Cross Section and Panel Data, 1st ed. Cambridge: MIT Press.

Chapter 2
Fixed Effects Models and Fixed Coefficients Models

Pietro Balestra and Jayalakshmi Krishnakumar

As noted in the introductory chapter, the simplest and most intuitive way to account for individual and/or time differences in behaviour, in the context of a panel data regression problem, is to assume that some of the regression coefficients are allowed to vary across individuals and/or through time. The regression coefficients are unknown, but fixed parameters. When these are allowed to vary in one or two dimensions, we speak of a *fixed effects model* (or fixed coefficients model).

It is useful, in this context, to distinguish between two types of regression coefficients: the intercept and the slope parameters. When only variations in the intercept are considered, the resulting regression problem is called a *covariance model* (or dummy variable model). Among the early proponents of such models in economics, one can cite Mundlak (1961), Hoch (1962), Kuh (1963) and Nerlove (1965). This model is discussed at length in Sect. 2.1 (for the case in which only individual variations occur) and in Sect. 2.2 (for the case in which both individual and time variations appear). The hypothesis of spherical disturbances which is typically maintained in such models is abandoned in Sect. 2.3, where different variance–covariance structure of the residuals are considered. Particular attention is paid to the problems of serial correlation and heteroscedasticity. Finally, some extensions (including variations of the slope parameters) are taken up in the last Section.

Pietro Balestra
Faculty of Economics, University of Lugano, via Giuseppe Buffi 13, CH-6904 Lugano, Switzerland

Jayalakshmi Krishnakumar
Department of Econometrics, University of Geneva, 40 Bd. Du pont d'Ave, CH-1211 Geneva 4, Switzerland, e-mail: jaya.krishnakumar@metri.unige.ch

Jaya Krishnakumar would like to dedicate this chapter to the memory of her dear teacher, colleague and friend Pietro Balestra who is no more. She feels really privileged to be the co-author of his last work.

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

2.1 The Covariance Model: Individual Effects Only

2.1.1 Specification

In this model, the intercept is allowed to vary from individual to individual, while the slope parameters are assumed to be constant in both the individual and time dimensions.

Consider, by contrast, the case in which all the parameters, including the intercept, are constant. Given a panel sample of N individuals over T periods, the basic linear regression equation takes the form:

$$\begin{aligned} y_{it} &= \alpha_0 + \beta_1 x_{1it} + \ldots + \beta_K x_{Kit} + \varepsilon_{it},\ i = 1,\ldots,N;\ t = 1,\ldots,T \\ &= \alpha_0 + x'_{it}\beta + \varepsilon_{it} \end{aligned} \tag{2.1}$$

where y_{it} is the observation on the dependent variable (for individual i at time t), x'_{it} is the K row vector of the explanatory variables, ε_{it} is a non-observable random term, β is the K column vector of the slope parameters and α_0 is the intercept.

When different intercepts are permitted for the N individuals, the model becomes:

$$y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it} \tag{2.2}$$

which is the basic (individual effect only) covariance model. The term covariance model is used with reference to the standard analysis of variance layout, which does not consider explicitly any explanatory variables. When the standard analysis of variance effects are combined with those of explanatory variables, the term covariance model is used.

Let us write down the model for the full sample. First, the T observations for individual i can be expressed conveniently in the following matrix form:

$$y_i = e_T \alpha_i + X_i \beta + \varepsilon_i \tag{2.3}$$

where y_i is the $T \times 1$ vector of the y_{it}, e_T is the unit vector of size T, X_i is the $T \times K$ matrix whose t-th row is x'_{it} and ε_i is the $T \times 1$ vector of errors. Next, stacking the individuals one after the other, we have:

$$\underset{(NT\times 1)}{\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_N \end{bmatrix}} = \underset{(NT\times N)}{\begin{bmatrix} e_T & 0 & \ldots & 0 \\ 0 & e_T & \ldots & 0 \\ & \ldots & & \\ 0 & 0 & \ldots & e_T \end{bmatrix}} \underset{(N\times 1)}{\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \ldots \\ \alpha_N \end{bmatrix}} + \underset{(NT\times K)}{\begin{bmatrix} X_1 \\ X_2 \\ \ldots \\ X_N \end{bmatrix}} \beta + \underset{(NT\times 1)}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_N \end{bmatrix}}$$

or more simply:

$$y = D_N \alpha + X\beta + \varepsilon\ . \tag{2.4}$$

The matrix D_N contains a set of N individual dummies, and has the following Kronecker product representation:

$$D_N = I_N \otimes e_T \ .$$

It can easily be verified that the following properties hold:

(i) $D_N e_N = e_N \otimes e_T = e_{NT}$ (exhaustivity)
(ii) $D_N' D_N = T I_N$ (orthogonality)
(iii) $D_N D_N' = I_N \otimes e_T e_T' = I_N \otimes J_T$
(iv) $\frac{1}{T} D_N' y = [\bar{y}_1, \ldots, \bar{y}_N]'$, $\frac{1}{T} D_N' X = [\bar{x}_1, \ldots, \bar{x}_N]'$

where $\bar{y}_i = \frac{1}{T} \sum_t y_{it}$ is the individual mean, $\bar{x}_i' = \frac{1}{T} \sum_t x_{it}'$ is the $K \times 1$ vector of the individual means of the explanatory variables and, by definition, $J_T = e_T e_T'$ is the unit matrix of order T.

Expression (2.4) represents the basic covariance model (in the case of individual effects only). To complete its specification, we adopt the following set of assumptions:

A1: The explanatory variables are non-stochastic, independent of the errors, and such that the $NT \times (N+K)$ matrix $Z = [D_N X]$ has full column rank;

A2: The random terms ε_{it} are independent, homoscedastic (variance σ^2) with zero mean.

Note that assumption A1 implies $NT > N + K$ (which is satisfied, for large N whenever $T \geq 2$), but also requires that the columns of X be linearly independent from those of D_N. For this to be the case, the matrices X_i must not contain the constant term (an obvious restriction) nor a column proportional to it (which precludes any variable, such as years of schooling, that is constant for a given adult individual, although varying from individual to individual).

2.1.2 Estimation

Given assumptions A1 and A2, the OLS estimators of all the regression coefficients in model (2.4) are BLUE. Collecting the regression coefficients in the vector γ, $\gamma' = [\alpha' \beta']$, and all the explanatory variables (including the N dummies) in the matrix $Z = [D_N X]$, the OLS estimator is $\hat{\gamma} = (Z'Z)^{-1} Z' y$. The actual computation of $\hat{\gamma}$ requires the inversion of the $(N+K) \times (N+K)$ matrix $Z'Z$, which for large N, is not an attractive operation. Instead, the technique of partitioned regression can be used, which involves the inversion of just a $(K \times K)$ matrix.

Using the results of partitioned regression or equivalently those of the Frisch–Waugh–Lovell theorem, these estimators (and other relevant statistics) can be expressed as:

$$\widehat{\beta} = (X' W_N X)^{-1} X' W_N y \tag{2.5}$$

$$\hat{\alpha} = (D_N' D_N)^{-1} D_N' (y - X\hat{\beta}) = \frac{1}{T} D_N' (y - X\hat{\beta}) \tag{2.6}$$

$$SS = y'W_N y - \hat{\beta}' X' W_N y \tag{2.7}$$

$$\hat{\sigma}^2 = SS/(NT - N - K) \tag{2.8}$$

$$V(\hat{\beta}) = \sigma^2 (X' W_N X)^{-1} \tag{2.9}$$

$$V(\hat{\alpha}) = \frac{\sigma^2}{T} I_N + \frac{1}{T} D_N' X V(\hat{\beta}) X' D_N \frac{1}{T} \tag{2.10}$$

where $W_N = I_{NT} - D_N (D_N' D_N)^{-1} D_N' = I_{NT} - \frac{1}{T} D_N D_N' = I_{NT} - I_N \otimes \frac{1}{T} J_T$ is an idempotent matrix of order NT and rank $NT - N$, also called the *within* projector (see the Appendix to this chapter).

The estimator given in (2.5) and the corresponding SS in (2.7) can also be obtained by OLS on the transformed model

$$y^* = X^* \beta + \varepsilon^* , \tag{2.11}$$

where $y^* = W_N y$ and $X^* = W_N X$. The transformation W_N is very simple: the transformed variables are simply the original variables expressed as *deviations from the individual mean* (the details are given is the Appendix). Therefore, the it-th equation corresponding to (2.11) is:

$$\begin{aligned} (y_{it} - \bar{y}_i) &= \beta_1 (x_{1it} - \bar{x}_{1i}) + \ldots + \beta_K (x_{Kit} - \bar{x}_{Ki}) + \varepsilon_{it}^* \\ &= (x_{it}' - \bar{x}_i') \beta + \varepsilon_{it}^* \end{aligned} \tag{2.12}$$

However, it should be remembered that, when working with transformed variables, the actual number of degrees of freedom is $NT - N - K$ and not $NT - K$ (as the above regression wrongly suggests), since in order to transform the variables, the N individual means must be computed, resulting in the loss of N degrees of freedom. Hence, when using a computer program on the transformed data, the variances given by the program must be adjusted accordingly.

If the transformed model is used, the estimators of the α_i and their respective variances and covariances can be obtained from (2.6) and (2.10), i.e.:

$$\begin{aligned} \hat{\alpha}_i &= \bar{y}_i - \bar{x}_i' \hat{\beta} \\ V(\hat{\alpha}_i) &= \frac{1}{T} \sigma^2 + \bar{x}_i' V(\hat{\beta}) \bar{x}_i \\ \mathrm{Cov}(\hat{\alpha}_i, \hat{\alpha}_j) &= \bar{x}_i' V(\hat{\beta}) \bar{x}_j \end{aligned}$$

There are other ways to eliminate the individual effects, but, when properly handled, they give rise to exactly the same estimator obtained previously, namely $\hat{\beta}$. Take the case of the popular first-difference transformation:

$$y_{it}^* = y_{it} - y_{i,t-1} \qquad t = 2, \ldots, T$$

or the less known deviation from the mean of the preceding observations:

$$y_{it}^{*} = y_{it} - \frac{1}{t-1}\sum_{s=1}^{t-1} y_{is} \qquad t = 2,\ldots,T$$

These two examples are special cases of the transformation $y_i^{*} = A'y_i$, $y^{*} = (I_N \otimes A')y$, where A' is a $(T-1)\times T$ matrix of full rank orthogonal to $e_T : A'e_T = 0$. For the two examples, the matrix A' takes the following form (respectively):

$$A' = \begin{bmatrix} -1 & 1 & & \\ & -1 & 1 & \\ \cdots & & & \\ & & -1 & 1 \end{bmatrix} \quad \text{and} \quad A' = \begin{bmatrix} -1 & 1 & & \\ -1/2 & -1/2 & 1 & \\ & \cdots & & \\ -\frac{1}{T-1} & -\frac{1}{T-1} & -\frac{1}{T-1} & \cdots\ 1 \end{bmatrix}$$

These transformations seem to offer an advantage, since they do not require any adjustment for degrees of freedom (the actual number of observations on the transformed variables available for estimation being the required $(NT-N)$. However, they introduce serial correlation (as in the first difference transformation) or heteroscedasticity (as in the second example) in the transformed model:

$$y^{*} = X^{*}\beta + \varepsilon^{*} \qquad V(\varepsilon^{*}) = \sigma^2(I_N \otimes A'A) = \sigma^2 V\ .$$

Hence OLS estimation is no longer BLUE.

Fortunately, the situation is one in which the pure GLS estimator is applicable (the variance–covariance matrix of ε^{*} being known up to a scalar multiple). This estimator is given by:

$$\begin{aligned} \hat{\beta}_{\mathrm{GLS}} &= (X^{*'}V^{-1}X^{*})^{-1}X^{*'}V^{-1}y^{*} \\ &= \left[X'(I_N \otimes A(A'A)^{-1}A')X\right]^{-1}X'(I_N \otimes A(A'A)^{-1}A')y \end{aligned}$$

In the Appendix it is shown that for any A such that $A'e_T = 0$, $A(A'A)^{-1}A' = I_T - \frac{1}{T}e_T e_T'$, and consequently $I_N \otimes A(\mathrm{A'A})^{-1}\mathrm{A'} = W_N$. The pure GLS estimator is thus equal to $\hat{\beta}$.

Actually the usual covariance estimator also belongs to this class, with A' composed of orthogonal rows (as in the second example above) but with unit length, so that $A'A = I_{T-1}$ and $V(\varepsilon^{*}) = \sigma^2 I_{N(T-1)}$. Consequently OLS can be applied directly to the transformed model and no correction for degrees of freedom is required. Such a matrix A' is obviously not unique. But if A' is chosen as in the second example above but with normalized rows, the transformation is given by:

$$y_{it}^{*} = \sqrt{\frac{t-1}{t}}\left(y_{it} - \frac{1}{t-1}\sum_{s=1}^{t-i} y_{is}\right) \qquad t = 2,\ldots,T$$

This particular matrix A' is called by Arellano (1995) the backward orthogonal deviations operator. Upon reverting the order of both rows and columns of A', one obtains the forward orthogonal deviations operator defined by:

$$y_{it}^{*} = \sqrt{\frac{T-t}{T-t+1}}\left(y_{it} - \frac{1}{T-t}\sum_{s=t+1}^{T} y_{is}\right) \qquad t = 1,\ldots,T-1$$

2.1.2.1 Consistency

Given the double dimension of panel data, asymptotic behavior can be studied in three different ways:

- Case 1: N fixed, $T \longrightarrow \infty$;
- Case 2: T fixed, $N \longrightarrow \infty$;
- Case 3: $N, T \longrightarrow \infty$.

The appropriate choice depends on the nature of the problem. For instance, if the N individuals refer to geographical region (i.e. the States in the U.S.), Case 1 is clearly indicated. The same would be true if N represents the number of the different industrial sectors in a given economy. However, if the individuals are a random sample from a large population (as is often the case in panel data models, with N large and T quite small), the relevant asymptotic is depicted in Case 2.

In all cases, the covariance estimator of the slope parameters (also called the within estimator) is consistent, under the usual regularity conditions. However, for the intercepts the situation is quite different. When N grows, the number of parameters α_i to be estimated becomes larger and larger. Therefore, the α_i can be estimated consistently only when N is fixed and T goes to infinity (Case 1). If, however, the true situation is the one corresponding to Case 2, the consistency problem of the individual effects can be circumvented by assuming that they are random variables rather than fixed parameters (see the chapter on error components models).

2.1.3 Inference

Under the normality assumption, the usual t-tests and F-tests can be performed. In particular, if one wishes to test the hypothesis $\alpha_i = \alpha_j$ (for some i and j, $i \neq j$) against the alternative $\alpha_i \neq \alpha_j$, the quantity

$$(\widehat{\alpha}_i - \widehat{\alpha}_j)/\sqrt{\widehat{V}(\widehat{\alpha}_i - \widehat{\alpha}_j)}$$

is distributed, under the null hypothesis, as a t-variable with $NT - N - K$ degrees of freedom.

An interesting question can be asked: are there any individual effects at all? The null hypothesis (of no individual effects) is in this case

$$\alpha_1 = \alpha_2 = \ldots = \alpha_N = \alpha_0$$

and the corresponding model (called the constrained model) is (2.1) with $NT - (K+1)$ degrees of freedom. Let SS_c be the sum of squared residuals of this

constrained model. The unconstrained model is (2.4) or, equivalently (2.11), with $NT-N-K$ degrees of freedom (and SS the corresponding sum of squares). Then, the quantity

$$\frac{(SS_c - SS)/(N-1)}{SS/(NT-N-K)}$$

is distributed as an F-variable with $(N-1)$ and $(NT-N-K)$ degrees of freedom. An F-test is therefore appropriate. The number of degrees of freedom in the numerator above is $N-1$ and not N, since testing that the N coefficients α_i are all equal is the same as testing that the $N-1$ differences $\alpha_{i+1}-\alpha_i$, $i=1,\ldots,N-1$, are all zero.

2.2 The Covariance Model: Individual and Time Effects

2.2.1 Time Effects Only

The treatment of time effects is analogous to that of individual effects. It would actually be the same if the observations were ordered by time first and then by individuals. However, it should be kept in mind that we maintain here the same ordering of the observations as in the preceding section.

Briefly stated, the time effect model is

$$y_{it} = \lambda_t + x'_{it}\beta + \varepsilon_{it}, \tag{2.13}$$

where to avoid confusion (and to permit the combination of time effects with individual effects) we use the symbol λ to designate a time varying intercept. For the full sample, the model becomes:

$$y = D_T\lambda + X\beta + \varepsilon \tag{2.14}$$

where D_T is a $NT \times T$ matrix of time dummies and λ is the $T \times 1$ vector of varying intercepts. In this case, the transformed model is a deviation from the time-means model (with $NT-T-K$ degrees of freedom), i.e.:

$$(y_{it} - \bar{y}_t) = (x'_{it} - \bar{x}'_t)\beta + \varepsilon^*_{it} \tag{2.15}$$

where $\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it}$ (the time-mean). Estimation and inference proceed as in the case of individual effects.

Written in full, the set of time dummies is:

$$D_T = \begin{bmatrix} I_T \\ I_T \\ \cdots \\ I_T \end{bmatrix} = e_N \otimes I_T$$

with the following properties:

(i) $D_T e_T = e_N \otimes e_T = e_{NT}$
(ii) $D_T' D_T = N I_T$
(iii) $D_T D_T' = e_N e_N' \otimes I_T = J_N \otimes I_T$
(iv) $\frac{1}{N} D_T' y = [\bar{y}_1, \ldots, \bar{y}_T]'$.

Finally, the deviation from the time-means matrix, called W_T, is:

$$W_T = I_{NT} - D_T (D_T' D_T)^{-1} D_T' = I_{NT} - \frac{1}{N} D_T D_T' = I_{NT} - \frac{1}{N} (J_N \otimes I_T) .$$

2.2.2 Time and Individual Effects

It would seem natural, in order to allow for generic individual and time effects, to include in the regression equation a full set of individual dummies and a full set of time dummies. However, this way of proceeding raises an identification problem. Not all the N coefficients α_i and the T coefficients λ_t are identifiable, since the columns of the matrix $[D_N D_T]$ are perfectly collinear (the sum of the first columns, $D_N e_N = e_{NT}$, is exactly equal to the sum of the last T columns, $D_T e_T = e_{NT}$).

It can be shown that the rank of $[D_N D_T]$ is exactly equal to $N+T-1$. Therefore, for identification, one column must be disregarded. This can be done arbitrarily, without affecting the result. Any set of $(N+T-1)$ linearly independent combinations of the columns of $[D_N D_T]$ will do. However, for the sake of symmetry, we prefer to delete one individual dummy (say the last) and one time dummy (again the last) and add the overall constant. Calling D_{N-1} and D_{T-1} the sets of $N-1$, and $T-1$ dummies respectively and α_* and λ_* the associated vectors of coefficients, the saturated individual and time effects covariance model can be written as:

$$y = e_{NT} c + D_{N-1} \alpha_* + D_{T-1} \lambda_* + X\beta + \varepsilon . \tag{2.16}$$

The value taken by the intercept for observation (i,t) can be easily read from the Table 2.1:

For the OLS estimation of (2.16), the Assumption A1 of the preceding section must be amended in the sense that now the matrix

$$[e_{NT}\ D_{N-1}\ D_{T-1}\ X]$$

Table 2.1 Values of the intercept

t \ i	$i < N$	$i = N$
$t < T$	$c + \alpha_{*i} + \lambda_{*t}$	$c + \lambda_{*t}$
$t = T$	$c + \alpha_{*i}$	c

must be of full column rank (equal to $N+T-1+K$). For this to be the case, the matrix X must not (as before) contain variables that assume constant values for each individual, nor (and this is peculiar to time effect models) admit variables that take a constant value for each time period (like, for instance, a price variable in a demand equation which is the same, at time t, for all individuals). Assumption A2 in obviously maintained.

Collecting all the dummies (including the overall constant) in the matrix $D = [e_{NT}\ D_{N-1}\ D_{T-1}]$ and calling $\gamma' = [c\ \alpha_*'\ \lambda_*']$ the vector of associated parameters, (2.16) can be put in the following compact way:

$$y = D\gamma + X\beta + \varepsilon \tag{2.17}$$

which is formally analogous to (2.4) of the individual effect model (with just D and γ replacing D_N and α). Using again the properties of partitioned regression, the following results hold:

$$\hat{\beta} = (X'W_{NT}X)^{-1}X'W_{NT}y \tag{2.18}$$

$$\hat{\gamma} = (D'^{-1}D'(y - X\hat{\beta})) \tag{2.19}$$

$$SS = y'W_{NT}y - \hat{\beta}'X'W_{NT}y \tag{2.20}$$

$$\hat{\sigma}^2 = SS/(NT - N - T + 1 - K) \tag{2.21}$$

$$V(\hat{\beta}) = \sigma^2(X'W_{NT}X)^{-1} \tag{2.22}$$

$$V(\hat{\gamma}) = \sigma^2(D'D)^{-1} + (D'D)^{-1}D'XV(\hat{\beta})X'D(D'D)^{-1} \tag{2.23}$$

where $W_{NT} = I_{NT} - D(D'D)^{-1}D'$ is an idempotent matrix of order NT and rank $NT - N - T + 1$, also called the *within* projector (for both individual and time effects).

Alternatively, one can apply OLS to the model transformed by W_{NT}, i.e.

$$y^* = X^*\beta + \varepsilon^* \tag{2.24}$$

where $y^* = W_{NT}y$ and $X^* = W_{NT}X$. What does this transformation represent? Can it be given as easy an interpretation as in the simple individual effects model? The answer is yes, but the algebra involved is somewhat complicated and better left for the Appendix. There it is shown that W_{NT} is composed of four separate transformations:

$$W_{NT} = I_{NT} - I_N \otimes \frac{1}{T}J_T - \frac{1}{N}J_N \otimes I_T + \frac{1}{NT}J_{NT} \tag{2.25}$$

the first being the identity transformation, the second the individual-mean transformation, the third the time-mean transformation and the last the overall-mean transformation. Therefore, the within transformation consists in subtracting from the original variables both the individual and time-means and in adding the overall-mean. The transformed model, accordingly, has the following simple expression:

$$(y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}) = (x_{it}' - \bar{x}_i' - \bar{x}_t' + \bar{x}_{it}'^*) \tag{2.26}$$

If the main interest lies in the slope parameters, then the transformed model should be used. It involves the inversion of a $K \times K$ matrix, while the dummy variable specification requires the inversion of a matrix of order $N+T-1+K$. Remember, however, that the exact number of degrees of freedom is $NT-N-T+1-K$ and not just $NT-K$ (requiring an adjustment for degrees of freedom when a regression package is used on the transformed data). From the transformed model, using (2.19) and (2.23), one can retrieve the coefficients c, α_* and λ_*, as shown below. But the computational cost may be too high. If one needs to estimate the intercept parameters (and their variances), it might be wise to use dummy variables.

For the sake of completeness (omitting the tedious details), the estimators of the various intercept parameters are given hereafter:

$$\begin{aligned}
\hat{c} &= (\bar{y}_N + \bar{y}_T - \bar{y}) - (\bar{x}_N' + \bar{x}_T' - \bar{x}')\hat{\beta} \\
\hat{\alpha}_{*i} &= (\bar{y}_i - \bar{y}_N) - (\bar{x}_i' - \bar{x}_N')\hat{\beta} \qquad i = 1, \ldots, N-1 \\
\hat{\lambda}_{*t} &= (\bar{y}_t - \bar{y}_T) - (\bar{x}_t' - \bar{x}_T')\hat{\beta} \qquad t = 1, \ldots, T-1
\end{aligned}$$

These results offer a neat interpretation that sheds some light on the identification problem. The parameter α_{*i} measures the fixed effect of individual i as deviation from the omitted individual effect; similarly, λ_{*t} measures the fixed effect of period t as deviation from the omitted time dummy; finally, the two omitted individual and time effects are included in the constant c, as deviation from the overall-mean. It is also possible to get exact analytical expressions for all variances and covariances, but, given their limited practical usefulness, they are not reported here.

It should be stressed, as a final point, that all the covariance estimators, given Assumptions A1 and A2, are BLUE. The consistency properties are similar to the ones discussed in Sect. 2.1.2. We thus have:

- Case 1: N fixed, $T \longrightarrow \infty$: only α_i's can be estimated consistently;
- Case 2: T fixed, $N \longrightarrow \infty$: only λ_t's can be estimated consistently;
- Case 3: $N, T \longrightarrow \infty$: a random effects approach is advisable as the number of parameters tends to infinity (see the next chapter).

The estimators of slope parameters are consistent in all the three cases.

2.2.3 Inference

In the present general covariance model, three interesting hypotheses may be tested:

– the absence of individual and time effects ($\alpha_* = 0$ and $\lambda_* = 0$):
– the absence of time effects (α_* free and $\lambda_* = 0$);
– the absence of individual effects ($\alpha_* = 0$ and λ_* free).

Assuming normality, these are F-tests. Denoting by:

– SS the sum of squares of the unrestricted covariance model ((2.17) or equivalently (2.26));

– SS_N the sum of squares of the individual dummy variables model ((2.4) or equivalently (2.12));
– SS_T the sum of squares of the time dummy model ((2.14) or equivalently (2.15));
– SS_c the sum of squares of the constant intercept regression model ((2.1));

the appropriate statistics for the three tests are, respectively:

$$F_1 = \frac{(SS_c - SS)/(N+T-2)}{SS/(NT-N-T+1-K)}$$

$$F_2 = \frac{(SS_N - SS)/(T-1)}{SS/(NT-N-T+1-K)}$$

$$F_3 = \frac{(SS_T - SS)/(N-1)}{SS/(NT-N-T+1-K)}$$

with degrees of freedom appearing, as deflators, in the numerator and denominator.

Before ending this section, let us mention that there is no particular problem regarding prediction in the fixed effect model and it is carried out in the usual manner taking into account the specific effect estimates.

2.3 Non-spherical Disturbances

2.3.1 What Variance–Covariance Stucture?

When the assumption of homoscedastic and independent errors (A2) is abandoned in favour of a more general variance–covariance structure for the residuals, the various estimators presented in the preceding pages are no longer BLUE (although they remain unbiased). Yet efficiency can be achieved by GLS or maximum likelihood techniques.

What is a typical variance–covariance structure in a panel data context? Two different types of correlation must now be considered: serial correlation, as in traditional time series analysis, and correlation among individuals. A general way of looking at this problem (given the adopted ordering of the observations) is to define the variance–covariance structure at the individual level, i.e.

$$E(\varepsilon_i \varepsilon_j') = A_{ij}\,,$$

where A_{ii} is the variance–covariance matrix of the errors of individual i (a positive definite matrix) and $A_{ij}, i \neq j$, is the covariance matrix between the residuals of individual i and individual j. Then, for the full sample, the variance–covariance matrix of ε takes the following form:

$$V(\varepsilon) = E(\varepsilon\,\varepsilon') = [A_{ij}] \qquad i,j = 1,\ldots,N\,.$$

with $A_{ji} = A_{ij}'$.

Several special cases are worth noting.

(a) **Serially Independent Errors**
In this case:

$$A_{ij} = \sigma_{ij} I_T \Rightarrow V(\varepsilon) = \Sigma \otimes I_T \ ,$$

where $\Sigma = [\sigma_{ij}]$ is the (constant) contemporaneous variance–covariance matrix. This specification was adopted by Zellner (1962) in his famous seemingly unrelated regression problem.

(b) **Individual Independence**
This case, characterized by

$$A_{ij} = 0, i \neq j \Rightarrow V(\varepsilon) = \text{diag}\,(\mathrm{A}_{11}, \ldots, \mathrm{A}_{\mathrm{NN}}) \ ,$$

covers all types of heteroscedasticity and serial correlation at the individual level.

(c) **Block-homoscedasticity**
It is an important special case of individual independence with in addition

$$A_{ii} = A \quad \forall\, i \Longrightarrow V(\varepsilon) = I_N \otimes A \ .$$

When $A = c_1 I_T + c_2 e_T e_T'$, where c_1 and c_2 are positive scalars, we have the (individual effect only) error component specification, studied in later chapters. In Sect. 2.3.3 below, we analyze in some details the case of serial correlation. (A being the variance–covariance matrix of a stationary stochastic process).

(d) **Block-equicorrelation**
It is defined by

$$\begin{cases} A_{ii} = A \ \forall\, i \\ A_{ij} = B \ i \neq j \end{cases} \Longrightarrow V(\varepsilon) = I_N \otimes (A - B) + e_N e_N' \otimes B$$

with A positive definite, B non negative definite and such that $A - B$ is positive definite. The special situation:

$$\begin{aligned} A &= (c_1 + c_3) I_T + c_2 e_T e_T' \\ B &= c_3 I_T \end{aligned} \tag{2.27}$$

(with c_1, c_2, c_3 positive scalars) corresponds to the full (both individual and time effects) error component specification.

2.3.2 Two General Propositions for Fixed Effects Models

Let us now go back to the basic covariance model, which we write as

$$y = L\gamma + X\beta + \varepsilon \tag{2.28}$$

where L represents a set of fixed effect dummies, either D_N for individual effects only), or D_T (for time effects only) or D (for both effects) and γ is the vector of associated parameters. In such a model, the choice of an appropriate variance–covariance structure naturally depends on the type of problem considered and very little can be said a priori. However, two general properties can be established, concerning on one hand the relevance of the within transformation and, on the other, the efficiency of the within estimator.

(a) **The Within Transformation**
In the case of spherical disturbances, the BLUE of β can be obtained directly by OLS on the transformed model

$$y^* = X^*\beta + \varepsilon^* \tag{2.29}$$

where $y^* = Wy$ and $X^* = WX$ with $W = I - L(L'^{-1}L')$. The simplicity of this transformation and its numerical advantages have been pointed out before. Yet, do these benefits carry over in a more general context characterised by a non-spherical variance structure? In other words, is it possible to apply first the within transformation and then use GLS? The answer is contained in the following proposition:

Proposition 2.1. *The GLS estimator of β in* (2.28) *is numerically equivalent to the GLS estimator of β in* (2.29) *using the same variance-covariance matrix $V(\varepsilon)$, for any observable X and y, if and only if $V(\varepsilon)L = LC$, for some non-singular C.*

For the proof of this proposition (and Proposition 2.2 below) see Aigner and Balestra (1988, Appendix).

It is easy to verify that the condition of Proposition 2.1 is satisfied in two interesting cases: (i) when $L = D_N$ and $V(\varepsilon) = \Sigma \otimes I_T$ (since $(\Sigma \otimes I_T)D_N = \Sigma \otimes e_T = (I_N \otimes e_T)\Sigma = D_N\Sigma$) and (ii) when $L = D_T$ with a block-equicorrelated covariance structure (since $[I_N \otimes (A-B) + e_N e'_N \otimes B]D_T = e_N \otimes (A-B) + Ne_N \otimes B = e_N \otimes (A + (N-1)B) = D_T(A + (N-1)B)$).

(b) **The Efficiency of the Within Estimator**
The following proposition holds.

Proposition 2.2. *In the presence of fixed effects, the usual within estimator of β is BLUE if and only if $V(\varepsilon)W = \lambda W$ for some positive scalar λ.*

An important situation in which the condition of this proposition is met occurs when the variance–covariance structure is of the error-component type (see the next chapter). Suppose that only individual effects are considered, i.e., $L = D_N$ and $V(\varepsilon) = I_N \otimes A$, $A = c_1 I_T + c_2 e_T e'_T$. Since W in this case is equal to W_N, we successively obtain:

$$\begin{aligned} V(\varepsilon)W_N &= I_N \otimes A - I_N \otimes \tfrac{1}{T} A e_T e'_T \\ &= I_N \otimes A(I_T - \tfrac{1}{T} e_T e'_T) \\ &= I_N \otimes c_1(I_T - \tfrac{1}{T} e_T e'_T) \\ &= c_1 W_N \end{aligned}$$

The reader can easily verify that the same is true when $L = D$ and the variance–covariance structure is of the general (both individual and time effects) error-component type.

An important conclusion emerges: fixed and random effects are two different alternative ways of considering heterogeneity in behaviour. They cannot be combined. The fixed effect model is particularly appropriate if we are interested in inferring on the behaviour of a specific set of N individuals and especially if the population consists only of these individuals (for example if the 'individuals' represent the different regions within a country and the sample covers all of them).

2.3.3 Individual Fixed Effects and Serial Correlation

Serial correlation is a common feature of time series data. How can it be handled in the context of a fixed effect model? Is it possible to recapture some of the computational advantages of the spherical case?

To answer these questions, let us reconsider the individual fixed effect model, (2.4), with serially correlated errors. The simplest possible scheme is an AR(1) process defined by

$$\varepsilon_{it} = \rho \varepsilon_{it-1} + u_{it}, \quad |\rho| < 1$$

with u_{it} a pure White Noise (with variance σ^2). Notice, and this is a crucial feature, that the parameters ρ and σ^2 are assumed to be the same for all individuals. In such a case, $V(\varepsilon_i) = \sigma^2 A$, with A a $(T \times T)$ matrix with typical element $a_{st} = \rho^{|t-s|}/(1-\rho^2)$. Furthermore, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$, and consequently $V(\varepsilon) = \sigma^2(I_N \otimes A)$. It will be recognised that the variance–covariance matrix of this problem is characterised by block-homoscedasticity. Therefore, the estimating procedure outlined below is valid not only for the AR(1) process but also for any ARMA process (with constant parameters across individuals) and indeed for any positive definite matrix A.

In this general setting, the BLUE of β can be obtained in two easy steps.

The first step consists in reducing the dimensionality of the problem by eliminating the individual dummies. As shown in Proposition 2.1, the usual *within* transformation works only under very special conditions, which are not satisfied here. But another easy transformation is available which works in all cases. It suffices to express each variable (both the dependent variable and each explanatory variable) as *deviation from the weighted individual mean*. For individual i, the *weighted individual mean* of the dependent variable is defined as

$$\bar{y}_i^w = e_T' A^{-1} y_i / e_T' A^{-1} e_T \tag{2.30}$$

and similarly for each explanatory variable. We shall denote by $(\bar{x}_i^w)'$ the $1 \times K$ vector of the weighted means of these variables.

The second step is to set up the generalized regression problem

$$\tilde{y}_i = \tilde{X}_i\beta + \tilde{\varepsilon}_i \Longleftrightarrow \tilde{y} = \tilde{X}\beta + \tilde{\varepsilon} \tag{2.31}$$

where $\tilde{y}_i$ and $\tilde{X}_i$ are the variables expressed as deviations from the weighted individual mean and treat it as if the variance–covariance matrix of the errors is the same as that of the original problem (i.e. $\sigma^2(I_N \otimes A)$). Two estimation procedures are available:

– either the GLS formula is used, leading to the estimator

$$\hat{\beta} = \left(\sum X_i' A^{-1} \tilde{X}_i\right)^{-1} \sum X_i' A^{-1} \tilde{y}_i$$

– or OLS is applied, after having transformed $\tilde{y}_i$ and $\tilde{X}_i$ by $P, (y_i^* = P\tilde{y}_i, \tilde{X}_i^* = P\tilde{X}_i)$ for P such that $PAP' = I$.

This two-step procedure also provides at no extra cost the BLUE estimators of the individual intercepts. These are simply given by:

$$\hat{\alpha}_i = \bar{y}_i^w - (\bar{x}_i^w)' \hat{\beta}$$

As an illustration, for the AR(1) process, the weighted individual mean is given by:

$$\frac{e_T' A^{-1} y_i}{e_T' A^{-1} e_T} = \frac{y_{i1} + y_{iT} + (1-\rho) \sum_{t=2}^{T-1} y_{it}}{2 + (T-2)(1-\rho)}$$

and P is the known Prais/Winsten transformation leading to:

$$\begin{aligned} y_{it}^* &= \sqrt{1-\rho^2}\tilde{y}_{it} \qquad t = 1 \\ &= \tilde{y}_{it} - \rho \tilde{y}_{it-1} \qquad t \geq 2 \end{aligned}$$

The above result for the AR(1) process is derived in a different manner by Bhargava Franzini and Narendranathan (1982). These authors also propose an adaptation of the Durbin–Watson serial correlation test for the individual effect model and discuss the problem of the estimation of the parameter ρ.

For the general case in which A is an unknown positive difinite matrix, a feasible GLS procedure can be implemented. The model is first estimated by the standard covariance method (or within transformation). From the computed residuals $\hat{\varepsilon}_i$, a consistent estimator of A (when $N \longrightarrow \infty$) is given by:

$$\hat{A} = \frac{1}{N} \sum \hat{\varepsilon}_i \hat{\varepsilon}_i' \, .$$

2.3.4 Heteroscedasticity in Fixed Effects Models

The two most common cases of departure from spherical disturbances are serial correlation and heteroscedasticity. We have just seen how to handle the first one in an appropriate manner. In this section we will see how heteroscedasticity across observations can be specified and dealt with in the context of a fixed effect model.

In particular, the availability of panel data will enable us to estimate the variances for some simple specifications of heteroscedasticity and thus apply the feasible GLS procedure. However, in a general setting, only corrections à la White/Newey-West can be recommended for want of adequate estimators for the variances parameters.

We will successively look at several cases going from a simple formulation to more complicated ones.

(a) **Individual heteroscedasticity only**
Here we will assume different variances for different individuals, constant over time with zero covariances. Thus:

$$\begin{aligned} E(\varepsilon_{it}^2) &= \sigma_i^2 \quad \forall t; \quad i = 1, \ldots, N \\ E(\varepsilon_i \varepsilon_i') &= \sigma_i^2 I_T \equiv V_i \end{aligned}$$

and

$$E(\varepsilon\varepsilon') = \text{diag}(\text{V}_i) \equiv \triangle \otimes \text{I}_T$$

where $\triangle = \text{diag}\ (\sigma_i^2)$.

This is a special case in which the condition of Proposition 2.1 is satisfied. Hence one can apply feasible GLS on the within transformed equation using $\hat{\triangle} \otimes I_T$ whose elements can be consistently estimated by

$$\hat{\sigma}_i^2 = \frac{1}{T} \sum_t \hat{\varepsilon}_{it}^{*2}$$

with $\hat{\varepsilon}_{it}^* = y_{it}^* - x_{it}^{*\prime} \hat{\beta}_w$, the within transformed residual.

(b) **Time-wise heteroscedasticity only**
Here we assume that the variances are different from one time period to another but constant over individuals for a given time period. Again zero covariances are assumed between different individuals. Therefore

$$\begin{aligned} E(\varepsilon_{it}^2) &= \sigma_t^2 \quad \forall i; \quad t = 1, \ldots, T \\ E(\varepsilon_i \varepsilon_i') &= \text{diag}(\sigma_t^2) \equiv \Lambda \end{aligned}$$

and

$$E(\varepsilon\varepsilon') = I_N \otimes \Lambda$$

This is a special case of block homoscedasticity where the covariance matrix can be expressed as $I_N \otimes A$. Hence we can follow the procedure described in Sect. 2.3.3 by taking deviations from the weighted means and applying feasible GLS. Consistent estimation of variances is given by:

$$\hat{\sigma}_t^2 = \frac{1}{N}\sum_i \hat{\varepsilon}_{it}^{*2}$$

(c) **Both individual and time-wise heteroscedasticity**
This is a combination of (a) and (b), maintaining zero covariances among different observations. We have:

$$\begin{array}{lll} E(\varepsilon_{it}^2) & = \sigma_{it}^2, & i = 1,\ldots,N; \quad t = 1,\ldots,T \\ E(\varepsilon_{it}\varepsilon_{js}) = 0 & & i \neq j \quad \text{or} \quad t \neq s \quad \text{or both} \end{array}$$

In this case there is no general feasible GLS estimator possible. Only the usual OLS/within estimators are available whose variance can be consistently estimated by applying the White correction:

$$V(\hat{\beta}_w) = (X^{*\prime}X^*)^{-1}\left(\sum_i\sum_t x_{it}^*\hat{\varepsilon}_{it}^{*2}x_{it}^{*\prime}\right)(X^{*\prime}X^*)^{-1}$$

If a specific structure is suspected for the occurrence of heteroscedasticity, say

$$\sigma_{it}^2 = h(a_0 + a_1 z_{1it} + \ldots + a_p z_{\text{pit}})$$

with $z_1,\ldots,z_p$ being observed variables (which may or may not overlap with the $x's$), then the usual two step procedure can be followed. First, estimate the auxiliary regression by OLS:

$$h^{-1}(\hat{\varepsilon}_{it}^{*2}) = a_0 + a_1 z_{1it} + \ldots + a_p z_{\text{pit}} + v_{it}$$

and estimate σ_{it}^2 as

$$\hat{\sigma}_{it}^2 = h(\hat{a}_0 + \hat{a}_1 z_{1it} + \ldots + \hat{a}_p z_{\text{pit}})$$

Then perform GLS (weighted least squares) by running OLS on

$$\frac{y_{it}}{\hat{\sigma}_{it}} = \frac{\alpha_i}{\hat{\sigma}_{it}} + \frac{x_{it}'}{\hat{\sigma}_{it}}\beta + \frac{\varepsilon_{it}}{\hat{\sigma}_{it}}$$

(d) **General block heteroscedasticity**
Here individuals are assumed to be independent, each having its own variances–covariance matrix. Thus

$$E(\varepsilon_i\varepsilon_i') = A_{ii}$$

and

$$E(\varepsilon_i\varepsilon_j') = 0 \qquad i \neq j$$

As in (c) above, no feasible GLS procedure is available in this case. The only solution consists in correcting the variance of the within estimator as follows: (Newey-West type correction (cf. Arellano (1987))

$$V(\hat{\beta}_w) = (X^{*\prime}X^*)^{-1}\left(\sum_i X_i^*\hat{\varepsilon}_i^*\hat{\varepsilon}_i^{*\prime}X_i^{*\prime}\right)(X^{*\prime}X^*)^{-1}$$

where $\hat{\varepsilon}_i^* = y_i^* - X_i^*\hat{\beta}_w$ is the within residual vector of individual i.

2.4 Extensions

2.4.1 Constant Variables in One Dimension

The generic fixed individual effect considered in Sect. 2.1 may be the result of some factors (such as sex, years of schooling, race, etc.) which are constant through time for any individual but vary across individuals. If observations are available on such variables, we may wish to incorporate them explicitly in the regression equation. The model may thus be written as:

$$y_{it} = q_i'\delta + x_{it}'\beta + \varepsilon_{it} \tag{2.32}$$

where the row-vector q_i' now contains the observations on the variables which are constant for individual i, including the constant term, and δ si the associated vector of coefficients. We shall assume that there are m such variables, in addition to the constant. The vector x_{it}', as before, contains K explanatory variables, varying in both dimensions.

Collecting the T observations for individual i, we get

$$y_i = (q_i' \otimes e_T)\delta + X_i\beta + \varepsilon_i \tag{2.33}$$

and finally, stacking the N individuals, we obtain the full model

$$y = (Q \otimes e_T)\delta + X\beta + \varepsilon \tag{2.34}$$

where Q is the $N \times (m+1)$ matrix whose i-th row is q_i'.

Next, we note that the columns of $(Q \otimes e_T)$ are linear combinations of the columns of the matrix of individual dummies. In fact:

$$(Q \otimes e_T) = (I_N \otimes e_T)(Q \otimes 1) = D_N Q$$

From this, we draw the following conclusions:

(i) if $m+1 > N$, the parameter vector δ is not identifiable. The slope parameters β can still be estimated by the covariance method (using the within transformation) in an unbiased and consistent way;
(ii) if $m+1 = N$, the matrix Q is square. Assuming that it is non-singular, then:
   – the BLUE estimator of β in (2.34) is the covariance estimator;

– the BLUE estimation of δ in (2.34) is a linear non-singular transformation of $\hat{\alpha}$ (the coefficient vector of the N individual dummies), i.e.:

$$\hat{\alpha} = Q\hat{\delta} \qquad \Longleftrightarrow \qquad \hat{\delta} = Q^{-1}\hat{\alpha}$$

(iii) if $m+1 < N$, the covariance estimator of β is no longer BLUE (but it is still unbiased). This is so because the dummy variable model and model (2.34) are related by the definition $\alpha = Q\delta$. When $m+1 < N$, a total of $N-m-1$ restrictions are imposed on the vector α. Ignoring these restrictions on α is like estimating a model with some additional extraneous variables, which produces unbiased but inefficient estimates. One can test the validity of such restrictions by the usual F-test. The constrained model is, in this case, model (2.34), while the unconstrained model is the individual dummy variables model.

From the above discussion it in clear that when constant individual variables are explicitly introduced in the regression equation there is no room for dummy variables (at least in an easily interpretable way).

The same argument applies when the model is extended to include variables that vary in time, but that are constant for all individuals (such as prices). Consider the regression equation:

$$y_{it} = c + q_i^{*\prime}\delta^* + p_t'\eta + x_{it}'\beta + \varepsilon_{it} \tag{2.35}$$

where q_i^* is the same vector as q_i but without the constant (with m components) and p_t' is the row-vector of n variables (without the constant) that are the same for all individuals at time t. For the full sample we have:

$$y = e_{NT}c + (Q^* \otimes e_T)\delta^* + (e_N \otimes P)\eta + X\beta + \varepsilon \tag{2.36}$$

where P is the $T \times n$ matrix whose t-th row is p_t'. Again it can be shown that:

(i) for identification of δ^* and η the necessary order conditions are $m < N$ and $n < T$;
(ii) the covariance estimator of β is unbiased and consistent in all cases;
(iii) when model (2.36) is just identified $(m+n = T+N-2)$, the covariance estimator of β is BLUE;
(iv) when $m+n < T+N-2$, it is like imposing some restrictions on the coefficients of the dummy variables.

2.4.2 Variable Slope Coefficients

In the covariance model, only the intercepts are allowed to vary across individuals and/or through time, while the slope parameters are kept constant. However, there are situations in which the slope parameters themselves may exhibit a pattern of variation.

Consider, for instance, the case in which all regression coefficients are individual specific:

$$y_{it} = \alpha_i + x_{it}'\beta_i + \varepsilon_{it} = z_{it}'\gamma_i + \varepsilon_{it} \tag{2.37}$$

where x_{it}' si a $(1 \times K)$ vector of explanatory variables, $z_{it}' = [1\ x_{it}']$ and $\gamma_i' = [\alpha_i\ \beta_i']$. With the usual notation, for individual i we write

$$y_i = e_T\alpha_i + X_i\beta_i + \varepsilon_i = Z_i\gamma_i + \varepsilon_i \tag{2.38}$$

and, analogously for the full sample:

$$y = D_N\alpha + \tilde{X}\beta + \varepsilon = \tilde{Z}\gamma + \varepsilon \tag{2.39}$$

where $\alpha' = [\alpha_1, \ldots, \alpha_N], \beta' = [\beta_1', \ldots, \beta_N'], \gamma' = [\gamma_1', \ldots, \gamma_N']$ and $\tilde{X}$ and $\tilde{Z}$ are the following block-diagonal matrices:

$$\tilde{X} = \begin{bmatrix} X_1 & & 0 \\ & \ddots & \\ 0 & & X_N \end{bmatrix} \qquad \tilde{Z} = \begin{bmatrix} Z_1 & & 0 \\ & \ddots & \\ 0 & & Z_N \end{bmatrix}$$

of order, respectively, $NT \times NK$ and $NT \times N(K+1)$. This is a special case of a SUR model (cf. Zellner (1962)) with independent errors across observations and equations (individuals in this context).

Given assumption A2, the BLUE estimator of γ in (2.39) is the OLS estimator. Simple algebra shows that the OLS estimator for the full sample boils down to the OLS estimator of each individual regression, (2.38), separately. Calling SS_i the sum of squared residuals of the i-th regression problem, the total sum of squares for the full model, denoted by SS, is simply the sum of the SS_i with $N(T-K-1)$ degrees of freedom. Note that the rank condition for the identification of the γ_i is that each matrix Z_i be of full column rank $K+1$ (which requires that $T > (K+1)$).

A test of homogeneity in behavior (all γ_i being equal to a common vector γ_0) can now be easily performed. The constrained model is the pooled model (with constant coefficients), i.e.,

$$y = Z\gamma_0 + \varepsilon \tag{2.40}$$

where $Z' = [Z_1', \ldots, Z_N']$. Its sum of squares is denoted by SS_c (with $NT-K-1$ degrees of freedom). Then, under normality, the following quantity

$$\frac{(SS_c - SS)/(N-1)(K+1)}{SS/N(T-K-1)}$$

is distributed as an F-variable with $(N-1)(K+1)$ and $N(T-K-1)$ degrees of freedom. An F-test is therefore appropriate.

When the model contains a constant term, an appropriate question to be asked is whether all the slope parameters are constant. In such a case, all intercepts are allowed to vary freely across individuals and the constrained model is the individual dummy variable model, whose sum of squares is denoted by SS_N, with $NT-N-K$ degrees of freedom. Then the quantity

$$\frac{(SS_N - SS)/(N-1)K}{SS/N(T-K-1)}$$

is distributed as an F-variable with $(N-1)K$ and $N(T-K-1)$ degrees of freedom.

In a similar way we can treat the case of time-dependent slopes (which boils down to OLS estimation period by period). We just write down, for future reference, the corresponding equations (for $N > K+1$):

$$\begin{aligned} y_{it} &= z'_{it}\lambda_t + \varepsilon_{it} \\ y_i &= \tilde{\tilde{Z}}_i\lambda + \varepsilon_i \\ y &= \tilde{\tilde{Z}}\lambda + \varepsilon \end{aligned}$$

where

$$\lambda' = [\lambda'_1, \ldots, \lambda'_T]$$

$$\tilde{\tilde{Z}}_i = \begin{bmatrix} z'_{i1} & & \\ & \ddots & \\ & & z'_{iT} \end{bmatrix} \qquad T \times T(K+1)$$

$$\tilde{\tilde{Z}}' = \left[\tilde{\tilde{Z}}'_1 \ldots \tilde{\tilde{Z}}'_N\right]$$

The most general approach in the case of fixed effects is to consider both individual and time variations in the coefficients. It is very tempting to write the model as:

$$y = \tilde{Z}\gamma + \tilde{\tilde{Z}}\lambda + Z\delta + \varepsilon \tag{2.41}$$

with $Z' = [Z'_1, \ldots, Z'_N]$, but for the same reason as in the covariance model the matrix of explanatory variables $[\tilde{Z}\ \tilde{\tilde{Z}}\ Z]$ is not of full column rank. In fact it is easy to see that the rank of this $NT \times (K+1)(N+T+1)$ matrix is at most equal to $(K+1)(N+T-1)$, since $\tilde{Z}(e_N \otimes I_{K+1}) = Z$ and $\tilde{\tilde{Z}}(e_T \otimes I_{K+1}) = Z$. Therefore $2(K+1)$ restrictions (at least) must be imposed on the regression coefficients. There are many different ways to do this.

One can, as Hsiao (1986) suggests, minimize the sum of squares in (2.41) subject to the $2(K+1)$ restrictions

$$\sum_{i=1}^{N} \gamma_i = 0, \quad \sum_{t=1}^{T} \lambda_t = 0\,,$$

or more simply (as was done in the covariance model), one could eliminate one vector γ_i (say the last, and the corresponding $K+1$ last columns of $\tilde{Z}$) and one vector λ_t (again the last, and the corresponding $K+1$ last columns of $\tilde{\tilde{Z}}$) and apply OLS. This amounts to using directly the two above restrictions in model (2.41) together with a slight reparameterisation of the coefficients. In all cases, however, the necessary order condition for identification $(NT > (K+1)(N+T-1))$ must be met.

2.4.3 Unbalanced Panels

Up to this point, our attention has focused entirely on genuine (or balanced) panels, a situation in which N individuals are observed over the same T time periods. When the number of observations is not the same for all individuals we speak of an unbalanced panel. In the present subsection we briefly discuss the incidence of unbalancedness on the estimation of fixed effect models.

When only individual fixed effects are considered, no particular new problem arises. The model can still be represented as in (2.4), except that now the matrix D_N of individual dummies does not have a nice Kronecker-product representation. Nonetheless, using the standard results of partitioned regression, it can easily be established that the deviation from the individual mean transformation (as in the balanced case) is the right transformation to be used in order to obtain the BLUE of β.

The story is quite different when both individual and time effects are deemed important. The difficulty stems from the fact that in this context the date of any single observation matters. Although it is always possible to work out the transformation that eliminates all fixed effect (the interested reader may consult in this respect the article by Wansbeek and Kapteyn(1989)), the procedure is too complicated for practical purposes. By far the easiest approach to this problem is to set up a regression equation containing, in addition to the explanatory variables, the overall constant $N-1$ individual dummies and T^*-1 time dummies (T^* being the set of all dates available in the sample).

Appendix: Matrix Algebra for Balanced Panels

In this Appendix we collect the basic orthogonal projectors which appear in panel data analysis and highlight their relationships and properties.

To make this expository note self confined (and useful for future reference) we recall the following notations:

e_s is the unit (column) vector of order s (whose elements are all equal to unity);
$J_s = e_s e_s'$ is the unit matrix of order $s \times s$;
y_{it} is the observation on a relevant variable for individual i $(i = 1, \ldots, N)$ at time $t (t = 1, \ldots, T)$;
y_i is the $(T \times 1)$ vector of observations for individual i;
y is the $(NT \times 1)$ vector of all observations : $y' = [y_1', \ldots, y_N']$;
$\overline{y} = \frac{1}{NT} \sum_i \sum_t y_{it}$ is the overall mean;
$\overline{y}_i = \frac{1}{T} \sum_t y_{it}$ is the individual mean;
$\overline{y}_t = \frac{1}{N} \sum_i y_{it}$ is the time mean.

The Basic Projectors

We consider a linear transformation of the vector $y : y^* = My$, where M is one of the orthogonal projectors commonly used in panel data analysis. The matrix M is idempotent (and symmetric) and can be viewed as the matrix of the idempotent quadratic form $y'My = y^{*\prime}y^*$.

We distinguish four cases.

Case 1: Just the overall effect

(1.a) The overall mean transformation: $y^*_{it} = \overline{y}$. It replaces each observation with the overall mean. The associated orthogonal projector is:

$$B_{NT} = \tfrac{1}{NT} e_{NT} e'_{NT} = \tfrac{1}{NT} J_{NT}$$

$$= \tfrac{1}{N} e_N e'_N \otimes \tfrac{1}{T} e_T e'_T = \tfrac{1}{N} J_N \otimes \tfrac{1}{T} J_T \qquad (\text{rank}\, B_{NT} = 1)$$

(1.b) The deviation from the overall mean transformation: $y^*_{it} = y_{it} - \overline{y}$. Associated projector: $DB_{NT} = I_{NT} - B_{NT}$ (rank = $NT - 1$)

Case 2: Individual effects only

(2.a) The individual mean transformation, also called (individual) *between* transformation: $y^*_{it} = \overline{y}_i$
Associated projector: $B_N = I_N \otimes \frac{1}{T} e_T e'_T = I_N \otimes \frac{1}{T} J_T$ (rank $= N$)

(2.b) The deviation from the individual mean transformation, also called (individual) *within* transformation: $y^*_{it} = y_{it} - \overline{y}_i$
Associated projector: $W_N = I_{NT} - B_N$ (rank $= NT - N$)

Case 3: Time effects only

(3.a) The time mean transformation, also called (time) *between* transformation: $y^*_{it} = \overline{y}_t$
Associated projector: $B_T = \frac{1}{N} e_N e'_N \otimes I_T = \frac{1}{N} J_N \otimes I_T$ (rank $= T$)

(3.b) The deviation from the time mean transformation, also called (time) *within* transformation: $y^*_{it} = y_{it} - \overline{y}_t$
Associated projector: $W_T = I_{NT} - B_T$ (rank = $NT - T$)

Case 4: Both individual and time effects

The (overall) *within* transformation: $y^*_{it} = y_{it} - \overline{y}_i - \overline{y}_t + \overline{y}$
Associated projector: $W_{NT} = I_{NT} - B_N - B_T + B_{NT}$ (rank $= (N-1)(T-1)$)

To check that each projector defined above does indeed perform the right transformation can be done easily using the selection vector $s'_{it} = (E^N_i)' \otimes (E^T_t)'$, where E^N_i is the i-th elementary vector of order N (and similarly for E^T_t). When applied to y, s'_{it} selects the observation $y_{it} : s'_{it} y = y_{it}$. We illustrate the procedure for the projector W_{NT} (from which all other cases can easily be derived). We obtain successively:

$$
\begin{aligned}
y_{it}^{*} &= s_{it}' y^{*} = s_{it}' W_{NT} y = s_{it}' y - s_{it}' B_N y - s_{it}' B_T y + s_{it}' B_{NT} y \\
&= y_{it} - \left[(E_i^N)' \otimes \frac{1}{T} e_T' \right] y - \left[\left(\frac{1}{N} e_i^N \right)' \otimes (E_T^t)' \right] y + \left[\frac{1}{N} e_N' \otimes \frac{1}{T} e_T' \right] y \\
&= y_{it} - \frac{1}{T} e_T' y_i - \frac{1}{N} \sum_i (E_T^t)' y_i + \frac{1}{NT} e_{NT}' y \\
&= y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}
\end{aligned}
$$

The Within Transformations and Dummy Variables

It is straightforward to verify that W_N and W_T are related to their respective sets of dummy variables in the following way:

- $W_N = I - D_N (D_N' D_N)^{-1} D_N'$ $\quad D_N = I_N \otimes e_T$ $\quad N$ individual dummies
- $W_T = I - D_T (D_T' D_T)^{-1} D_T'$ $\quad D_T = e_N \otimes I_T$ $\quad T$ time dummies

For W_{NT} the situation is more delicate. We cannot use the full set of $N+T$ dummy variables, since the matrix $[D_N \; D_T]$ is not of full rank. This can be seen by noting that the sum of the first N columns $(D_N e_N = e_{NT})$ is equal to the sum of the last T columns $(D_T e_T = e_{NT})$. Therefore we have to choose just $N+T-1$ linearly independent columns of $[D_N \; D_T]$ or any non-singular transformation of them (the result being invariant to any non singular transformation). To establish the result in the simplest way, we choose to keep the full set of time dummies (D_T) and add $N-1$ linearly independent combinations of the columns of D_N, i.e. $D_* = D_N A$, with A an $N \times (N-1)$ matrix of full rank. Since D_* must also be linearly independent when associated with D_T, the combination to avoid is $D_N e_N$. To ensure this, we choose, in A, $(N-1)$ columns orthogonal to $e_N : A' e_N = 0$. We then define the matrix

$$D = [D_* \; D_T]$$

Now we observe that the columns of D_* are orthogonal to those of D_T, $D_*' D_T = A' D_N' D_T = A'(e_N \otimes e_T') = A' e_N e_T' = 0$, so that:

$$
\begin{aligned}
D(D'D)^{-1} D' &= D_* (D_*' D_*)^{-1} D_*' + D_T (D_T' D_T)^{-1} D_T' \\
&= D_N A (A' D_N' D_N A)^{-1} A' D_N' + B_T \\
&= \frac{1}{T} D_N A (A'A)^{-1} A' D_N' + B_T
\end{aligned}
$$

Next we note that $e_N (e_N' e_N)^{-1} e_N' + A(A'^{-1} A' = I_N$ (since the matrix $F = [e_N A]$ is non singular and therefore $I = F(F'F)^{-1} F' = e_N (e_N' e_N)^{-1} e_N' + A(A'A)^{-1} A'$ and consequently

$$D(D'D)^{-1}D' = \frac{1}{T}D_n D_N' - \frac{1}{NT}D_N e_N e_N' D_N' + B_T$$
$$= B_N - \frac{1}{NT}e_{NT}e_{NT}' + B_T = B_N - B_{NT} + B_T$$

This establishes that:

$$W_{NT} = I - D(D'D)^{-1}D'$$

A final remark: the three within projectors W_N, W_T, W_{NT} have the property of eliminating all constant effects of the appropriate type (individual, time, or both).

Relationships Between the Different Projectors

For *Cases 1 to 3*, we have defined for each case two matrices, call them M_1 and M_2, which have the following three properties:

- they are idempotent (and symmetric): $M_i M_i = M_i \quad i = 1,2$;
- they are mutually orthogonal: $M_i M_j = 0 \quad i \neq j$;
- their sum is the identity matrix: $M_1 + M_2 = I$.

For *Case 4*, the situation requises some careful attention. We do indeed have four idempotent matrices $(W_{NT}, B_T, B_N, B_{NT})$, but they do not share the last two properties. In order to get a decomposition fulfilling the three properties, we rewrite W_{NT} as:

$$W_{NT} = I_{NT} - (B_N - B_{NT}) - (B_T - B_{NT}) - B_{NT}$$

and define the four idempotent matrices:

$M_1 = B_N - B_{NT}$ (deviation of the individual mean from the overall mean)
$M_2 = B_T - B_{NT}$ (deviation of the time mean from the overall mean)
$M_3 = B_{NT}$ (the overall mean)
$M_4 = W_{NT}$ (the overall within transformation)

It is now easy to verify that :

$M_i M_i = M_i \quad$ all i
$M_i M_j = 0 \quad i \neq j$
$\sum M_i = I$.

Properties of the Decomposition

Suppose that s idempotent matrices of order n, M_i $i = 1, \ldots, s$, satisfy the above three conditions and define the following positive definite matrix:

$$A = \sum_i \lambda_i M_i \qquad \lambda_i > 0 \qquad \lambda_i \neq \lambda_j$$

Then:

1. The scalars λ_i are eigenvalues of the matrix A with multiplicity $r_i = \text{rank}\,(M_i)$
2. $\mid A \mid = \Pi_i \lambda_i^{r_i}$
3. $A^{-1} = \sum_i \lambda_i^{-1} M_i$
4. $QAQ = I$ for $Q = \sum_i \lambda_i^{-1/2} M_i$.

To prove these results is extremely simple. For result (1), it suffices to post-multiply A successively by $M_1, M_2, \ldots$. Since $AM_i = \lambda_i M_i$, the columns of M_i are eigenvectors of A associated with λ_i. Given that there are r_i linearly independent columns in M_i (ad noting that $\sum r_i = n$), the multiplicity of λ_i is exactly equal to r_i. Result (2) follows from the fact that the determinant of a matrix is equal to the product of its eigenvalues. Finally, result (3) and (4) are verified by simple multiplication ($AA^{-1} = I$ and $QAQ = I$).

The usefulness of these results is that the variance–covariance matrix of standard error component models can be expressed in the form of matrix A.

References

Aigner, D.J. and P. Balestra (1988), "Optimal Experimental Design for Error Components Models", *Econometrica,* 56, (4), 955–971.

Arellano, M. (1987), "Computing Robust Standard Errors for Within-Groups Estimators", *Oxford Bulletin of Economics and Statistics*, 49, 431–434.

Arellano, M. and O. Bover (1995), "Another Look at the Instrumental Variables Estimation of Error Component Models", *Journal of Econometrics*, 68, 29–51.

Bhargava, A., L. Franzini and W. Narendranathan (1982), "Serial Correlation and Fixed Effects Model", *Review of Economic Studies*, 49, 533–549.

Hoch, I. (1962), "Estimation of Production Function Parameters Combining Time Series and Cross Section Data", *Econometrica*, 30, 34–53.

Hsiao, C. (1986), "Analysis of Panel Data", *Cambridge University Press*, Cambridge.

Kuh, E. (1963), "The Validity of Cross Sectionally Estimated Behaviour Equations in Time Series Applications", *Econometrics*, 27, 197–214.

Mundlak, Y. (1961), *Capital Stock Growth: A Micro-Econometric Approach*, Amsterdam: North-Holland.

Nerlove, M. (1965), *Estimation and Indentification of Cobb-Dougkas Production Functions*, Chicago: Rand McNally.

Wansbeek, T. and A. Kapteyn (1989), "Estimation of the Error Components Model with Incomplete Panels", *Journal of Econometrics*, 41, 341–361.

Zellner, A. (1962), "An Efficient Method of Estimating Seemingly Unrelated Regression and Test for Aggregation Bias", *Journal of the American Statistical Association*, 57, 348–368.

Chapter 3
Error Components Models

Badi H. Baltagi, László Mátyás and Patrick Sevestre

3.1 Introduction

As discussed in the previous chapters, the disturbances of an econometric model include all factors affecting the behavior/phenomenon under study that the econometrician cannot explicitly specify, because the relevant statistical information either does not exist or is not accessible. This is the so-called unobserved heterogeneity. As an example, factors such as personal ability, adaptability, work diligence, etc. do have an impact on employees' wage profile but are generally not observed (see Chap. 22 on this). As long as they can be assumed not to vary over time, they can be accounted for through individual effects. Such individual effects also allow us to account for unobserved factors affecting, for example, firms behavior (regarding their investments in general and their foreign direct investment in particular, their labor demand, and/or their production efficiency; see Chaps. 20 and 21 below). Indeed, firms' environment as well as their managers' behavior impact upon their efficiency and employment/investment decisions whilst they are not fully observed.

As already stressed in the first chapters, one of the major advantages of panel data is that their "double dimension" enables us to account for these unobservable factors as long as they can be considered fixed over time. The main difference between the fixed effects models considered in the previous chapter and the error components models considered here is the assumption made about those individual effects.

Badi H. Baltagi
Center for Policy Research, 426 Eggers Hall, Syracuse University, Syracause, NY 13244-1020, USA, e-mail: bbaltagi@maxwell.syr.edu

László Mátyás
Central European University, Department of Economics, Nádor u. 9, 1051 Budapest, Hungary, e-mail: matyas@ceu.hu

Patrick Sevestre
Université Paris 1 – Panthéon Sorbonne, Ecole Economique de Paris (Paris School of Economics), 106-112 Boulevard de l'Hôpital, 75013 Paris, France, e-mail: Patrick.Sevestre@univ-paris1.fr

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

The basic assumption underlying the error components model is the absence of correlation between the individual effects and the regressors of the model. Although this (quite restrictive) assumption can be relaxed (see Chap. 4 in particular), we shall stick to it in this chapter as the error components model can be considered as one of the pillars of panel data econometrics.

3.2 The One-Way Error Components Model

3.2.1 Definition/Assumptions of the Model

The one-way error components model can be written as:

$$y_{it} = \beta_0 + \sum_{k=1}^{K} \beta_k x_{it} + u_{it} \ , \quad i = 1, \ldots, N \ \text{ and } \ t = 1, \ldots, T \tag{3.1}$$

with

$$u_{it} = \alpha_i + \varepsilon_{it} \ .$$

The disturbances u_{it} are decomposed into two components, α_i and ε_{it}, which explains the model's name. α_i represents the individual effects, accounting for unobservable factors affecting y and which do not vary over time; ε_{it} represents the other variables influencing y but which vary both over time and individuals. Both are assumed to be independently distributed across individuals.[1] Another important assumption underlying the error components model is the strict exogeneity of regressors, which implies:

$$E(\alpha_i \mid x_{i1}, x_{i2}, \ldots, x_{iT}) = 0, \ \ \forall i$$
$$E(\varepsilon_{it} \mid x_{i1}, x_{i2}, \ldots, x_{iT}) = 0, \ \ \forall i, t$$

Moreover, α_i and ε_{it} are both assumed to be serially uncorrelated and homoskedastic:

$$V(\alpha_i \mid x_{i1}, x_{i2}, \ldots, x_{iT}) = \sigma_\alpha^2, \ \ \forall i$$
$$\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{it'} \mid x_{i1}, x_{i2}, \ldots, x_{iT}) = \delta_{tt'}\sigma_\varepsilon^2, \ \ \forall i, t, t'$$
$$\mathrm{Cov}(\alpha_i, \varepsilon_{it} \mid x_{i1}, x_{i2}, \ldots, x_{iT}) = 0, \ \ \forall i, t.$$

Given the assumption stating the absence of correlation between the regressors and the individual effects, the latter do not affect the conditional expectation of y but do impact its variance. Indeed, given the above assumptions, one has:

$$E(y_{it} \mid x_{i1}, x_{i2}, \ldots, x_{iT}) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{it}$$

[1] See Chap. 18 for a specific framework in which such a correlation across individuals can (must) be assumed.

and

$$\begin{aligned}\mathrm{Cov}(y_{it}, y_{it'} \mid x_{i1}, x_{i2}, \ldots, x_{iT}) &= \mathrm{Cov}(u_{it}, u_{it'} \mid x_{i1}, x_{i2}, \ldots, x_{iT}) \\ &= \begin{cases} \sigma_\alpha^2 + \sigma_\varepsilon^2 & \text{if } t = t' \\ \sigma_\alpha^2 & \text{if } t \neq t' . \end{cases}\end{aligned}$$

The presence of the individual effects in the disturbances thus induces, for each individual, some serial correlation across periods. It is worthwhile noticing that this serial correlation does not depend on the time interval between two observations, contrary to the usual pattern of serial correlation in time-series models.

Stacking all the observations related to the individual i, one can write:

$$\underset{(T\times 1)}{y_i} = \underset{(T\times(k+1))}{X_i} \times \underset{((k+1)\times 1)}{\beta} + \underset{(T\times 1)}{u_i} \tag{3.2}$$

where $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$ represents the vector of observations of the dependent variable for the ith individual; X_i the matrix of observations of the explanatory variables (including the constant term) and u_i the vector of the disturbances for this individual. Given the assumptions defining this model, the vector of the disturbance terms has the following properties:

$$\begin{aligned} E(u_i \mid x_{i1}, x_{i2}, \ldots, x_{iT}) &= 0, \ \forall i \\ V(u_i \mid x_{i1}, x_{i2}, \ldots, x_{iT}) &= A, \ \forall i \end{aligned}$$

with:

$$\underset{(T\times T)}{A} = \begin{pmatrix} \sigma_\alpha^2 + \sigma_\varepsilon^2 & \sigma_\alpha^2 & \ldots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\varepsilon^2 & \ldots & \sigma_\alpha^2 \\ \ldots\ldots & . & \ldots & . \\ \ldots\ldots & . & \ldots & . \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \ldots & \sigma_\alpha^2 + \sigma_\varepsilon^2 \end{pmatrix}$$

$$= \sigma_\varepsilon^2 I_T + \sigma_\alpha^2 J_T$$

where I_T is the identity matrix of order T and J_T is a square $(T \times T)$ matrix of ones.

The pattern of this matrix clearly shows the existence of a serial correlation associated with the individual effects; magnitude of which is independent of the time span between the time periods under consideration.

Stacking then the whole set of individual vectors of observations, $y = (y_{11}, y_{12}, \ldots, y_{1T}, \ldots, y_{N1}, y_{N2}, \ldots, y_{NT})'$, such that the slower index is i and the faster index is t, one can write the model as:

$$\underset{(NT\times 1)}{y} = \underset{(NT\times(k+1))}{X} \times \underset{((k+1)\times 1)}{\beta} + \underset{(NT\times 1)}{u} \tag{3.3}$$

with:

$$E(u \mid x_1, x_2, \ldots, x_K) = 0$$
$$V(u \mid x_1, x_2, \ldots, x_K) = \sigma_\varepsilon^2\, \Omega$$

where:

$$\underset{(NT,NT)}{\sigma_\varepsilon^2\, \Omega} = \begin{pmatrix} E(\varepsilon_1 \varepsilon_1^{'}) & E(\varepsilon_1 \varepsilon_2^{'}) & \ldots & E(\varepsilon_1 \varepsilon_N^{'}) \\ E(\varepsilon_2 \varepsilon_1^{'}) & E(\varepsilon_2 \varepsilon_2^{'}) & \ldots & E\varepsilon_2 \varepsilon_N^{'}) \\ \ldots.. & . & \ldots & . \\ \ldots.. & . & \ldots & . \\ E(\varepsilon_N \varepsilon_1^{'}) & E(\varepsilon_N \varepsilon_2^{'}) & \ldots & E(\varepsilon_N \varepsilon_N^{'}) \end{pmatrix}$$

$$= \begin{pmatrix} A & 0 & \ldots & 0 \\ 0 & A & \ldots & 0 \\ \ldots.. & . & \ldots & . \\ \ldots.. & . & \ldots & . \\ 0 & 0 & \ldots & A \end{pmatrix}$$

$$= I_N \otimes A$$
$$= \sigma_\varepsilon^2\, [I_{NT} + (\sigma_\alpha^2/\sigma_\varepsilon^2)\, (I_N \otimes J_T)]$$

i.e., using the Within and Between matrix transforms notations:[2]

$$V(u) = \sigma_\varepsilon^2\, [W_N + ((\sigma_\varepsilon^2 + T\sigma_\alpha^2)/\sigma_\varepsilon^2) B_N]\,.$$

The nullity of all off-diagonal blocks in Ω just corresponds to the assumed independence across individuals.

The error components model then appears as a rather common regression model with a particular pattern of serial correlation in the disturbances. The well-known result that the OLS estimator of the coefficients is still unbiased and consistent but inefficient then applies. It is also much well-known that the GLS estimator has better properties in this context.

3.2.2 The GLS Estimator

3.2.2.1 Definition

Recall that in a model where the variance–covariance matrix of the disturbances is proportional to Ω, the GLS estimator of the coefficients is given by:

$$\widehat{\beta}_{\text{gls}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y.$$

[2] $W_N = I_{NT} - (I_N \otimes J_T/T)$ and $B_N = I_N \otimes J_T/T$.

Given the particular structure of Ω,[3] this estimator can be written as:

$$\widehat{\beta}_{\text{gls}} = (X'W_N X + \theta X'B_N X)^{-1}(X'W_N y + \theta X'B_N y)$$

where W_N and B_N are respectively the Within and Between operatorst and

$$\theta = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2} .$$

The GLS estimator then combines the Within and Between variation of the observations. It does so in an optimal way as it can be shown that GLS corresponds to the minimum variance linear unbiased estimator among all the estimators combining the Within and Between variation. In other words, the value of λ that minimizes the variance of:

$$\widehat{\beta}(\lambda) = (X'W_N X + \lambda X'B_N X)^{-1}(X'W_N\, y + \lambda X'B_N\, y)$$

is just equal to $\lambda = \sigma_\varepsilon^2/(\sigma_\varepsilon^2 + T\sigma_\alpha^2) = \theta$.

Then, if we knew the value of Ω (or, equivalently, that of θ), computing the GLS estimator would be very simple as it is well-known that GLS can be interpreted as OLS on the transformed model:

$$\Omega^{-1/2}\, y = \Omega^{-1/2}\, X\, \beta + \Omega^{-1/2}\, u\,,$$

which, in this particular case, just amounts to:

$$y_{it} + (\sqrt{\theta} - 1)\bar{y}_i = [x_{it}' + (\sqrt{\theta} - 1)\bar{x}_i'\,]\,\beta + \varepsilon_{it} + (\sqrt{\theta} - 1)\bar{\varepsilon}_i$$

where $\bar{y}_i = \sum_t y_{it}$, etc.

3.2.2.2 Properties

Given the assumptions stated above, the GLS estimator is unbiased and efficient. Its variance is given by:

$$\begin{aligned} V(\widehat{\beta}_{\text{gls}}) &= \sigma_\varepsilon^2 (X'\Omega^{-1}X)^{-1} \\ &= \sigma_\varepsilon^2 (X'W_N X + \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2} X'B_N X)^{-1} . \end{aligned}$$

Moreover, if the disturbances are normally distributed, the GLS estimator is also normally distributed:

[3] Because the spectral decomposition of Ω is given by $W_N + ((\sigma_\varepsilon^2 + T\sigma_\alpha^2)/\sigma_\varepsilon^2)B_N = W_N + (1/\theta)B_N$, one gets $\Omega^{-1} = W_N + \theta B_N$.

$$\widehat{\beta}_{\text{gls}} \sim N\ (\beta\,, \sigma_\varepsilon^2 (X'WX + \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2} X'BX)^{-1}).$$

This estimator is consistent for N going to infinity, with finite T, under the same assumptions as above[4] and if the variance–covariance matrices of the regressors are bounded both in the Between and Within dimensions, i.e., if:

$$\text{plim}_{N\to\infty} \frac{X'B_N X}{N} = B_{xx}^N\,, \text{ a finite positive definite matrix}$$

and,

$$\text{plim}_{N\to\infty} \frac{X'W_N X}{N} = W_{xx}^N\,, \text{ a finite positive definite matrix.}$$

If moreover, the terms α_i and ε_{it} in the disturbances are independently and identically distributed, the GLS estimator is asymptotically efficient and its asymptotic distribution (when $N \to \infty$ but T is fixed) is given by:

$$\sqrt{N}(\widehat{\beta}_{\text{gls}} - \beta) \sim N\,(0\,, \sigma_\varepsilon^2 (W_{xx}^N + \theta B_{xx}^N)^{-1})\,.$$

Along the same lines, and with assumptions adapted from the above ones, it is possible to show that the GLS estimator is also consistent when both N and T tend to infinity. Its asymptotic distribution is then given by:

$$\sqrt{NT}(\widehat{\beta}_{\text{gls}} - \beta) \sim N\,(0\,, \sigma_\varepsilon^2 (W_{xx}^{NT})^{-1})$$

with

$$\text{plim}_{N,T\to\infty} \frac{X'W_N X}{NT} = W_{xx}^{NT}, \text{ a finite positive definite matrix.}$$

Indeed, when T tends to infinity, θ tends to 0 and $\widehat{\beta}_{\text{gls}}$ converges to the Within estimator $\widehat{\beta}_w$ described below. It is then obvious that the GLS estimator's variance–covariance matrix does not depend on the Between variation of the regressors.

Thus, the GLS estimator exhibits good properties, both in finite samples and asymptotically. Unfortunately, it relies on unknown parameters, namely the variance of the individual effects σ_α^2 and that of the idiosyncratic element, σ_ε^2 that appear in $\theta = \sigma_\varepsilon^2/(\sigma_\varepsilon^2 + T\sigma_\alpha^2)$. In order to compute a "feasible-GLS" estimator, one then has to estimate first the variance components σ_α^2 and σ_ε^2 in order to get in turn an estimate of θ.

[4] More rigorously, the "Error components-GLS" estimator remains consistent even when one mis-specifies the variance–covariance matrix of the disturbances. If the true matrix is given by $V(u) = \sigma_\varepsilon^2 \Sigma \neq \sigma_\varepsilon^2 \Omega$, the GLS estimator is consistent as long as the matrix $X'\Omega^{-1}\Sigma\Omega^{-1}X\ /\ N$, converges towards a positive definite matrix. But the variance–covariance matrix of the GLS estimated coefficients is clearly inconsistent in this case.

3.2.3 The Feasible GLS Estimator

3.2.3.1 Definition

The Feasible-GLS estimator definition is very close to that of true GLS except for the fact that Ω is replaced by a consistent estimator $\widehat{\Omega}$:

$$\begin{aligned}\widehat{\beta}_{\text{fgls}} &= (X'\widehat{\Omega}^{-1}X)^{-1}X'\widehat{\Omega}^{-1}y \\ &= (X'W_N X + \widehat{\theta}X'B_N X)^{-1}(X'W_N\, y + \widehat{\theta}X'B_N\, y)\end{aligned}$$

with

$$\widehat{\theta} = \frac{\widehat{\sigma}_\varepsilon^2}{\widehat{\sigma}_\varepsilon^2 + T\widehat{\sigma}_\alpha^2}\ .$$

3.2.3.2 Variance Estimation

There are several ways to estimate the unknown variances σ_α^2 and σ_ε^2. However, the most commonly used approach is that proposed by Swamy and Arora (1972). It consists of using the residual variances associated with the regression stated in the Within and Between dimensions of the observations. Let us first consider the former, i.e., the regression based on the Within transformation of the equation:

$$y_{it} - \overline{y}_i = [x_{it}^{'} - \overline{x}_i^{'}]\beta + \varepsilon_{it} - \overline{\varepsilon}_i\ .$$

That is, in matrix form,

$$W_N\, \overline{y} = W_N\, X\, \beta + W_N\, \varepsilon\ .$$

In this model:

$$V(W_N\, \varepsilon) = \sigma_\varepsilon^2 W_N\ .$$

As we know from Kruskal's theorem, even though the variance–covariance matrix of the disturbances is not scalar, OLS on this model is still the BLUE of β. As a consequence, the natural estimator of the variance σ_ε^2 is given by:

$$\begin{aligned}\widehat{\sigma}_w^2 &= \frac{\widehat{\varepsilon}_w'\widehat{\varepsilon}_w}{\text{rank}(M_{wx})} \\ &= \frac{\widehat{\varepsilon}_w'\widehat{\varepsilon}_w}{N(T-1)-k_w}\end{aligned}$$

with

$$\begin{aligned}\widehat{\varepsilon}_w &= W_N\, y - W_N\, X\widehat{\beta}_w \\ &= (W_N\ - W_N\, X(X'W_N\, X)^{-1}X'W_N\ )y \\ &= (W_N\ - P_{wx})y \\ &= M_{wx}y\end{aligned}$$

which is exactly the residual from the Within regression.

It is quite easy to show that this estimator is unbiased. Indeed,

$$\begin{aligned}
E(\widehat{\varepsilon}\,' W_N \widehat{\varepsilon}) &= \text{trace}[E(\widehat{\varepsilon}_w' \widehat{\varepsilon}_w)] \\
&= \text{trace}[E(\varepsilon' M_{wx} \varepsilon)] \\
&= E[\text{trace}(M_{wx} \varepsilon \varepsilon')] \\
&= \text{trace}(M_{wx}).E(\varepsilon \varepsilon') \\
&= \text{trace}[M_{wx}.\sigma_\varepsilon^2 \Omega] \\
&= \sigma_\varepsilon^2 \times \text{trace}[M_{wx}(W_N + (1/\theta) B_N)] \\
&= \sigma_\varepsilon^2 \times \text{trace}[M_{wx}] \\
&= \sigma_\varepsilon^2 \times \text{trace}[W_N - P_{wx}] \\
&= \sigma_\varepsilon^2 \times (N(T-1) - k_w)
\end{aligned}$$

where k_w is the number of regressors in the Within regression. As a consequence,

$$\begin{aligned}
E(\widehat{\sigma}_w^2) &= E[\frac{\widehat{\varepsilon}_w' \widehat{\varepsilon}_w}{N(T-1) - k_w}] \\
&= \sigma_\varepsilon^2 \, .
\end{aligned}$$

It is worthwhile to note that caution must be exercised when computing this variance using software packages. Indeed, when estimating the Within regression with OLS on transformed data, the usual software packages will generally consider that there are NT observations and k_w estimated coefficients. The number of degrees of freedom considered in the computations of variances will then be $(NT - k_w)$. This is not correct as the true number of degrees of freedom should account for the individual effects that are implicitly estimated when doing a Within regression. The correct number of degrees of freedom should be $N(T-1) - k_w$ (i.e., the rank of the W_N operator).

Once this potential problem of degrees of freedom is correctly accounted for, the residual variance of the Within regressions provides a consistent estimator of σ_ε^2 (when $N \to \infty, T$ being finite as well as when N and $T \to \infty$).

Proceeding along the same lines, one can show that the residual variance of the Between regression, i.e. the variance resulting from applying OLS to the model:

$$B_N y = B_N X\beta + B_N u$$

allows to get a consistent estimate of $\sigma_\alpha^2 + \sigma_\varepsilon^2/T$. Indeed, let us consider the residuals of this regression: $\widehat{u}_b = B_N y - B_N X\widehat{\beta}_b = M_{bx} y$ (computed on NT observations), one can show that

$$\begin{aligned}
E(\widehat{\sigma}_b^2) &= E[\frac{\widehat{u}_b' \widehat{u}_b}{T(N - k_b)}] \\
&= \sigma_\alpha^2 + \sigma_\varepsilon^2/T.
\end{aligned}$$

Here again, one must pay attention to the degrees of freedom. Using NT observations, the software will generally set the number of degrees of freedom to $NT - k_b$ where k_b is the number of regressors in the Between regression while the correct number of degrees of freedom is equal to $T(N - k_b)$, due to the T times repetition of the individual means. On the other hand, if the Between regression is computed on the sample of the N individual means, i.e. if one estimates the model

$$\bar{y}_i = \bar{x}_i\beta + \bar{u}_i \qquad i = 1, \ldots, N$$

the residual variance then directly provides a consistent estimate of $\sigma_\alpha^2 + \sigma_\varepsilon^2/T$

$$\begin{aligned} E(\widehat{\sigma}_b^2) &= E[\frac{\sum_i \bar{\widehat{u}}_i^2}{N - k_b}] \\ &= \sigma_\alpha^2 + \sigma_\varepsilon^2/T \ . \end{aligned}$$

Then, a consistent estimate of θ can easily be computed as:

$$\widehat{\theta} = \frac{\widehat{\sigma}_w^2}{T\widehat{\sigma}_b^2} = \frac{\widehat{\sigma}_\varepsilon^2}{\widehat{\sigma_\varepsilon^2 + T\sigma_\alpha^2}} \ .$$

There are many other ways to estimate the variances.[5] However, Maddala and Mount (1973) have shown that the choice of a particular method to estimate these variances does not impact significantly on the properties of the estimated coefficients in the second step of the Feasible-GLS estimator (see also Taylor (1980)).

This does not mean that replacing the true value θ by an estimate $\widehat{\theta}$ does not have any consequence. Although it does not affect the asymptotic properties of the feasible GLS estimator, it does have some influence on its finite sample properties. Indeed, while the GLS estimator is unbiased, the Feasible GLS is not, except under very particular circumstances (see Taylor (1980)).

3.2.3.3 Properties

As just stated above, the Feasible-GLS estimator of the error components model is, as any other Feasible-GLS estimator, biased in finite samples

$$\begin{aligned} E(\widehat{\beta}_{\text{fgls}}) &= E[(X'\widehat{\Omega}^{-1}X)^{-1}X'\widehat{\Omega}^{-1}y] \\ &= \beta + E[(X'\widehat{\Omega}^{-1}X)^{-1}X'\widehat{\Omega}^{-1}u] \\ &\neq \beta \ . \end{aligned}$$

Indeed, the dependence between $\widehat{\Omega}$ and the disturbances u impairs the nullity of the second term in the above expectation and thus induces a bias for $\widehat{\beta}_{\text{fgls}}$. However,

[5] See Wallace and Hussain (1969), Amemiya (1971), Swamy and Arora (1972).

Taylor (1980) has shown that when the disturbances are normally distributed, $\widehat{\beta}_{\text{fgls}}$ is unbiased as long as the variances σ_α^2 and σ_ε^2 are estimated via the Between and Within regressions (cf. supra) and that $N \geq k+5$ and $T \geq 2$. Moreover, while we know that, in an econometric model with unknown heteroskedasticity or serial correlation, the true variance of the feasible-GLS estimator of the coefficients is, in general, unknown, Taylor (1980) has nevertheless provided the formula of the variance–covariance matrix of the Feasible-GLS estimator of the error components model assuming normality of the disturbances and $N \geq k+10$ and $T \geq 2$. Unfortunately, the expression of this variance is quite complex and its real computation not easy. However, an interesting by-product of this result is that Taylor (1980) has shown that this Feasible-GLS estimator is often more precise than the other estimators available for the error components model. But a further drawback of this estimator is that, in finite samples, it is not distributed as a Normal, even when the disturbances are.

The "unsatisfactory" finite sample properties of the feasible-GLS estimator are quite "classical". Also "classical" are its good asymptotic properties: it is asymptotically equivalent to the GLS estimator as long as N tends to infinity (whatever T). Then, under the error components assumptions and assuming that $X'\widehat{\Omega}^{-1}X/N$ converges to a finite positive definite matrix, the Feasible-GLS estimator is consistent for N tending to infinity, T finite.[6] Its asymptotic distribution is then given by

$$\sqrt{N}(\widehat{\beta}_{\text{fgls}} - \beta) \sim N(0, \sigma_\varepsilon^2 (B_{xx}^N + \theta W_{xx}^N)^{-1}) .$$

Indeed, the Feasible-GLS estimator is asymptotically equivalent to the GLS one. It is then asymptotically efficient. Those properties remain valid when both N and T go to infinity. In particular,

$$\sqrt{NT}(\widehat{\beta}_{\text{fgls}} - \beta) \sim N(0, \sigma_\varepsilon^2 (W_{xx}^{NT})^{-1}) .$$

Then, under the assumptions stated above (which include the strict exogeneity of all regressors), the Feasible-GLS estimator provides very reliable parameter estimates, at least if the individual dimension of the sample is large enough.

3.2.4 Some Other Estimators

GLS and Feasible-GLS thus combine in an optimal way the Within and Between variance of the observations. However, this does not preclude other ways to estimate an error components model. Indeed, comparing those different estimators to each other may be helpful in identifying possible mis-specification errors.

[6] cf. footnote 4.

3.2.4.1 The OLS Estimator

OLS on the pooled data yields:

$$\widehat{\beta}_{\text{ols}} = (X'X)^{-1}X'y\,. \tag{3.4}$$

Under the assumptions which we considered earlier in this chapter, this estimator is unbiased and its variance for any linear regression is given by

$$V(\widehat{\beta}_{\text{ols}}) = \sigma_\varepsilon^2 (X'X)^{-1}X'\Omega X(X'X)^{-1}\,.$$

When N tends to infinity but T remains finite, the OLS estimator is consistent. If the error components α_i and ε_{it} are i.i.d., OLS is asymptotically distributed as a Normal

$$\sqrt{N}(\widehat{\beta}_{\text{ols}} - \beta) \;\sim\; N\,(0\,,\, \sigma_\varepsilon^2 (B_{xx}^N + W_{xx}^N)^{-1}(W_{xx}^N + \theta B_{xx}^N)(B_{xx}^N + W_{xx}^N)^{-1})\,.$$

But one has to notice that while OLS is still consistent when both N and T go to infinity, its asymptotic variance is unbounded in this case (e.g. see Trognon (1993)).

3.2.4.2 The Between Estimator

This estimator just amounts to applying OLS to the model written in terms of individual means

$$\overline{y}_i = \sum_{k=1}^{K} \beta_k \overline{x}_{ki} + \overline{u}_i\,, \qquad \text{with} \;\; \overline{u}_i = \alpha_i + \overline{\varepsilon}_i, \quad i = 1, \ldots, N\,.$$

Using the above defined Between operator B_N, the model can be written, in matrix form, as,

$$\underset{(NT\times 1)}{B_N y} \;=\; \underset{(NT\times k_b)}{B_N X} \;\times\; \underset{(k_b\times 1)}{\beta} \;+\; \underset{(NT\times 1)}{B_N u}\,.$$

It is worth mentioning here that the latter way of writing the model induces T repetitions of the model written in the former way, i.e. in terms of the individual means. However, writing the model in such a matrix form allows a more systematic way of presenting the Between estimator. This is just given by:

$$\widehat{\beta}_B = (X'B_N X)^{-1}X'B_N y\,. \tag{3.5}$$

which means that the Between estimator makes use of the Between individual variation of the observations only, hence giving a full weight to "permanent" differences across individuals.

Under the assumptions stated above, this estimator is unbiased and its variance is

$$\begin{aligned} V(\widehat{\beta}_B) &= \sigma_\varepsilon^2 (X' B_N X)^{-1} X' B_N \, \Omega B_N \, X (X' B_N X)^{-1} \\ &= (\sigma_\varepsilon^2 + T\sigma_\alpha^2)(X' B_N X)^{-1}. \end{aligned}$$

Moreover, it is consistent when N goes to infinity, but T remains finite as well as when both N and T go to infinity. Its asymptotic distribution (for $N \to \infty$) is normal

$$\sqrt{N}(\widehat{\beta}_B - \beta) \;\sim\; N\,(0\,, \sigma_\alpha^2 (B_{xx}^N)^{-1})\,.$$

However, as the OLS estimator, the Between estimator does not have a finite variance when N and T go to infinity.

3.2.4.3 The Within Estimator

This estimator, also called "covariance estimator", or "fixed effects estimator", is OLS applied to the model written in differences from individual means

$$y_{it} - \overline{y}_i = (x_{it}' - \overline{x}_i')\beta + \varepsilon_{it} - \overline{\varepsilon}_i$$

In matrix form

$$\underset{(NT\times 1)}{W_N\, y} \;=\; \underset{(NT\times k_w)}{W_N\, X} \;\times\; \underset{(k_w\times 1)}{\beta} \;+\; \underset{(NT\times 1)}{W_N\, \varepsilon}\,.$$

The Within estimator then writes as

$$\widehat{\beta}_W = (X' W_N X)^{-1} X' W_N \, y\,. \tag{3.6}$$

This method thus makes use of the Within-individual variation of the observations only. This can be seen as a drawback which discards a large part of the information contained in the raw data. Indeed, in most panels, the differences across individuals is often larger than that "Within-individuals". Moreover, as the above formula clearly shows, this estimator is identical to the one obtained under the assumption that the individual effects are fixed, and not random. As a consequence, the constant term as well as the coefficients of the explanatory variables which are constant over time cannot be estimated using this method. Indeed, let us consider the model

$$y_{it} = X_{it}\beta + Z_i c + \varepsilon_{it}\;;$$

the Within estimator amounts to discarding the Z_i variables from the model, due to the "Within transformation":

$$y_{it} - \overline{y}_i = (x_{it}' - \overline{x}_i')\beta + \varepsilon_{it} - \overline{\varepsilon}_i\,.$$

One can solve this problem by applying OLS to

$$\overline{y}_i - \overline{x}_i' \widehat{\beta}_W = Z_i c + \overline{u}_i \, .$$

Under the assumptions considered since the beginning of this chapter, $\widehat{\beta}_W$ is unbiased and its variance is given by

$$\begin{aligned} V(\widehat{\beta}_W) &= (X'W_N X)^{-1} X'W_N \Omega W_N X (X'W_N X)^{-1} \\ &= \sigma_\varepsilon^2 (X'W_N X)^{-1} \, . \end{aligned}$$

This estimator is consistent, both when N goes to infinity, with finite T and when N and T simultaneously go to infinity. The corresponding asymptotic distributions are normal

$$\sqrt{N}(\widehat{\beta}_W - \beta) \; \sim \; N\,(0\,, \sigma_\varepsilon^2 (W_{xx}^N)^{-1})$$

and

$$\sqrt{NT}(\widehat{\beta}_W - \beta) \; \sim \; N\,(0\,, \sigma_\varepsilon^2 (W_{xx}^{NT})^{-1})) \, .$$

It is remarkable that when both N and T go to infinity, the asymptotic distribution of the Within estimator is identical to that of the GLS estimator (this is because $\lim_{T\to\infty} \theta = \lim_{T\to\infty} \sigma_\varepsilon^2 / (\sigma_\varepsilon^2 + T\sigma_\alpha^2) = 0$). As a consequence, the Within estimator is asymptotically efficient in this case.

3.2.4.4 Reinterpreting Usual Estimators: The λ-type Estimators

All the estimators considered until now make use of either the Between variance of the observations, their Within variance, or both. It is then quite natural to group those estimators together, within a class, that may be called "$\lambda - class$" estimators, defined as (see Maddala (1979)):

$$\begin{aligned} \widehat{\beta}(\lambda) &= [X'W_N X + \lambda X' B_N X]^{-1} (X'W_N y + \lambda X' B_N y) \\ &= [X'(W_N + \lambda B_N)X]^{-1} X'(W_N + \lambda B_N) y \end{aligned}$$

where λ is a scalar such that,

- if $\lambda = 0$, $\widehat{\beta}(\lambda) = \widehat{\beta}_W$; one gets the Within estimator;
- if $\lambda = \theta$, $\widehat{\beta}(\lambda) = \widehat{\beta}_{\text{gls}}$; one gets the GLS estimator;
- if $\lambda = \widehat{\theta}$, $\widehat{\beta}(\lambda) = \widehat{\beta}_{\text{fgls}}$; one gets the Feasible-GLS estimator;
- if $\lambda = 1$, $\widehat{\beta}(\lambda) = \widehat{\beta}_{\text{ols}}$; one gets the OLS estimator;
- if $\lambda = \infty$, $\widehat{\beta}(\lambda) = \widehat{\beta}_B$; one gets the Between estimator.

3.2.4.5 The Maximum Likelihood Estimator

Making the additional assumption that both the individual effects α_i and the idiosyncratic disturbances ε_{it} are normally distributed as $N(0, \sigma_\alpha^2)$ and $N(0, \sigma_\varepsilon^2)$ respectively, it is also possible to resort to the maximum likelihood principle to estimate the error components model. The log-likelihood attached to this model is given by

$$\begin{aligned}\ln(L) &= -\frac{NT}{2}\ln(2\pi) - \frac{NT}{2}\ln(\sigma_\varepsilon^2) + \frac{N}{2}\ln(\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}) \\ &\quad - \frac{1}{2\sigma_\varepsilon^2}(y - X\beta)'\Omega^{-1}(y - X\beta) \\ &= -\frac{NT}{2}\ln(2\pi) - \frac{NT}{2}\ln(\sigma_\varepsilon^2) + \frac{N}{2}\ln(\theta) \\ &\quad - \frac{1}{2\sigma_\varepsilon^2}(y - X\beta)'\Omega^{-1}(y - X\beta)\end{aligned}$$

with $\Omega^{-1} = W_N + \theta\, B_N$. Maximizing this log-likelihood with respect to β, σ_ε^2 and θ does not yield closed form expressions, given the non-linearity in θ. However, one can make use of the first order conditions with respect to β and σ_ε^2:

$$\frac{\partial \ln(L)}{\partial \beta} = -\frac{1}{\widehat{\sigma}_w^2} X'\Omega^{-1}(y - X\widehat{\beta}) = 0$$

$$\frac{\partial \ln(L)}{\partial \sigma_\varepsilon^2} = -\frac{NT}{2\widehat{\sigma}_\varepsilon^2} + \frac{1}{\widehat{\sigma}_\varepsilon^4}(y - X\widehat{\beta})'\Omega^{-1}(y - X\widehat{\beta}) = 0$$

in order to concentrate the likelihood. Indeed, from the above equations, we get

$$\widehat{\beta}_{\mathrm{MLE}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

and

$$\widehat{\sigma}^2_{\varepsilon \mathrm{MLE}} = \frac{1}{NT}(y - X\widehat{\beta}_{\mathrm{MLE}})'\Omega^{-1}(y - X\widehat{\beta}_{\mathrm{MLE}}) = 0\,.$$

Then, following Breusch (1987), one can concentrate the likelihood, i.e., substitute $\widehat{\beta}$ and $\widehat{\sigma}_\varepsilon^2$ for their true values β and σ_ε^2 so that the likelihood now only depend on one unknown parameter, θ:

$$\ln(L) = -\frac{NT}{2}(1 + \ln(2\pi)) - \frac{NT}{2}\ln[(y - X\widehat{\beta})'(W_N + \theta B_N)(y - X\widehat{\beta})] + \frac{N}{2}\ln(\theta)\,.$$

Conditionally on β, maximizing this log-likelihood with respect to θ leads to

$$\theta = \frac{(y - X\beta)'W(y - X\beta)}{(T-1)(y - X\beta)'B(y - X\beta)}\,.$$

Then, an iterative procedure can be set up: taking the Within estimator $\widehat{\beta}_W$ as the departure estimate of the procedure, one can estimate θ using the above formula by replacing $\widehat{\beta}_W$ for β and then going back to estimating β by $\widehat{\beta}_{\mathrm{MV}}$ as defined above. Breusch (1987) has shown that the sequence of the $\widehat{\theta}_{(j)}$'s obtained at each iteration (j) of the procedure forms a monotonic sequence, so that this procedure should lead to a global maximum.

Given the assumptions made, these estimators are consistent when only $N \to \infty$ and also when both N and $T \to \infty$. Moreover, $\widehat{\beta}_{\mathrm{MV}}$ has an asymptotic Normal distribution, identical to that of the GLS estimator (cf. Trognon (1993).

3.2.5 Prediction

Suppose we want to predict y for the ith individual, S periods ahead. For the model given in (3.3), knowing the variance–covariance structure of the disturbances, Goldberger (1962) showed that the best linear unbiased predictor (BLUP) of $y_{i,T+S}$ is

$$\widehat{y}_{i,T+S} = X'_{i,T+S}\widehat{\beta}_{\mathrm{gls}} + \varpi' \frac{\Omega^{-1}}{\sigma_\varepsilon^2}\widehat{u}_{\mathrm{gls}} \quad \text{for } s \geqslant 1$$

where $\widehat{u}_{\mathrm{gls}} = y - X\widehat{\beta}_{\mathrm{gls}}$ and $\varpi = E(u_{i,T+S}\, u)$. Note that for period $T+S$

$$u_{i,T+S} = \alpha_i + \varepsilon_{i,T+S}$$

and $\varpi = \sigma_\alpha^2(l_i \otimes e_T)$ where l_i is the ith column of I_N, i.e. l_i is a vector that has 1 in the ith position and zero elsewhere and e_T is defined in (2.3). In this case

$$\varpi' \frac{\Omega^{-1}}{\sigma_\varepsilon^2} = \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}(l_i' \otimes e_T')[W_N + (\sigma_\varepsilon^2/(\sigma_\varepsilon^2 + T\sigma_\alpha^2))B_N] = \frac{\sigma_\alpha^2}{(\sigma_\varepsilon^2 + T\sigma_\alpha^2)}(l_i' \otimes e_T')$$

since $(l_i' \otimes e_T')B_N = (l_i' \otimes e_T')$ and $(l_i' \otimes e_T')W_N = 0$. The typical element of $\varpi' \frac{\Omega^{-1}}{\sigma_\varepsilon^2}\widehat{u}_{\mathrm{gls}}$ becomes $(\frac{T\sigma_\alpha^2}{(\sigma_\varepsilon^2 + T\sigma_\alpha^2)})\overline{\widehat{u}}_{i,\mathrm{gls}}$ where $\overline{\widehat{u}}_{i,\mathrm{gls}} = \sum_{t=1}^{T} \widehat{u}_{it,gls}/T$.

Therefore, the BLUP for $y_{i,T+S}$ corrects the GLS prediction by a fraction of the mean of the GLS residuals corresponding to that ith individual (see Taub (1979)). Baillie and Baltagi (1999) consider the practical situation of prediction from the error components model when the variance components are not known. They derive both theoretical and simulation evidence as to the relative efficiency of four alternative predictors:

(i) an ordinary predictor, based on the optimal predictor given above, but with MLEs replacing population parameters,
(ii) a truncated predictor that ignores the error components correction, given by the last term above, but uses MLEs for its regression parameters,
(iii) a misspecified predictor which uses OLS estimates of the regression parameters, and
(iv) a fixed effects predictor which assumes that the individual effects are fixed parameters that can be estimated.

The asymptotic formula for MSE prediction are derived for all four predictors. Using numerical and simulation results, these are shown to perform adequately in realistic sample sizes. Both the analytical and sampling results show that there are

substantial gains in mean square error prediction by using the ordinary predictor instead of the misspecified or the truncated predictors, especially with increasing $\rho = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2)$ values. The reduction in MSE is about ten fold for $\rho = 0.9$ and a little more than two fold for $\rho = 0.6$ for various values of N and T. The fixed effects predictor performs remarkably well being a close second to the ordinary predictor for all experiments. Simulation evidence confirm the importance of taking into account the individual effects when making predictions. The ordinary predictor and the fixed effects predictor outperform the truncated and misspecified predictors and are recommended in practice.

3.3 More General Structures of the Disturbances

The previous model can be generalized in several ways, by allowing for more general types of serial correlation as well as for possible heteroskedasticity. Let us first consider the two-way error components model, i.e., the model with both individual and time specific effects in the disturbances.

3.3.1 The Two-Way Error Components Model

The two-way error components model allows for specific time effects (λ_t) accounting for unobserved factors assumed to affect all individuals in a similar way at a given point in time.

3.3.1.1 Definition/Assumptions of the Model

This model can be written as

$$y_{it} = x_{it}'\beta + u_{it}$$

with

$$u_{it} = \alpha_i + \lambda_t + \varepsilon_{it} \ ,$$

or, in vector form for all observations

$$y = X\beta + u$$

with

$$u = \alpha \otimes e_T + (e_N \otimes I_T)\lambda + \varepsilon$$

where λ is the random vector of time effects $(T \times 1)$.

As in the previous model we assume that α, λ and ε are mutually independent, with 0 means and variance–covariance matrices

$$E(\alpha\alpha') = \sigma_\alpha^2 I_N\,, \qquad E(\lambda\lambda') = \sigma_\lambda^2 I_T \qquad E(\varepsilon\varepsilon') = \sigma_\varepsilon^2 I_{NT}\,.$$

Since the individual and time effects are incorporated in the model through the error structure, our main interest has to focus, as earlier, on the covariance matrix of the disturbance terms

$$\begin{aligned} E(uu') &= \sigma_\alpha^2(I_N \otimes J_T) + \sigma_\lambda^2(J_N \otimes I_T) + \sigma_\varepsilon^2 I_{NT} \\ &= \sigma_\varepsilon^2 \Omega\,. \end{aligned}$$

3.3.1.2 The GLS Estimator

If we want to use the GLS estimator, as in the case of the one-way error components model, we need the inverse of the covariance matrix Ω. Starting from Ω and using its spectral (eigen value) decomposition

$$\Omega^{-1} = W_{NT} + \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2} B_N + \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + N\sigma_\lambda^2} B_T + \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2 + N\sigma_\lambda^2} \frac{J_{NT}}{NT}\,.$$

Now the GLS estimator is

$$\widehat{\beta}_{\text{gls}} = [X'(W_{NT} + \theta B_N + \theta_1 B_T + \theta_2 \frac{J_{NT}}{NT})X]^{-1} X'(W_{NT} + \theta B_N + \theta_1 B_T + \theta_2 \frac{J_{NT}}{NT})y$$

where

$$\begin{aligned} \theta &= \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}\,, \\ \theta_1 &= \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + N\sigma_\lambda^2}\,, \\ \theta_2 &= \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2 + N\sigma_\lambda^2}\,. \end{aligned}$$

It would seem that this estimators is not very operational. However, one can get GLS as an OLS regression by transforming the equation as follows:

$$\begin{aligned} &(W_{NT} + \sqrt{\theta} B_N + \sqrt{\theta}_1 B_T + \sqrt{\theta}_2 \frac{J_{NT}}{NT})y \\ &= [y_{it} - (1 - \sqrt{\theta})\bar{y}_i - (1 - \sqrt{\theta}_1)\bar{y}_t + (1 - \sqrt{\theta} - \sqrt{\theta}_1 + \sqrt{\theta}_2)\bar{y} \end{aligned}$$

The small sample properties of the GLS estimator in this model are clearly the same as for the model with only individual effects. So the GLS remains unbiased and BLUE. When only N goes to infinity but T stays finite, the GLS is also consistent and has an asymptotic distribution given by

$$\sqrt{N}(\widehat{\beta}_{\text{gls}} - \beta) \sim N(0, \sigma_\varepsilon^2(\widetilde{W}_{xx}^N + \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2} B_{xx}^N)^{-1}) ,$$

as long as

$$\lim_{N\to\infty} \frac{X'X}{N} = T_{xx}^N \text{ is a finite positive definite matrix}$$

as well as

$$\lim_{N\to\infty} \frac{1}{N} X' W_{NT} X = \widetilde{W}_{xx}^N ,$$

and where B_{xx}^N is defined as in the one-way error components model.

In the case where both N and $T \to \infty$, the GLS is consistent and its asymptotic distribution is

$$\sqrt{NT}(\widehat{\beta}_{\text{gls}} - \beta) \sim N(0, \sigma_\varepsilon^2 W_{xx}^{NT\,-1}) .$$

under the hypotheses

$$\lim_{N\&T\to\infty} \frac{1}{NT} X'X = T_{xx}^{NT}$$

is a finite positive definite matrix,[7] and

$$\lim_{N\&T\to\infty} \frac{1}{NT} X' W_{NT} X = \widetilde{W}_{xx}^{NT} \text{ is also positive definite.}$$

Despite its good properties, the GLS estimator is unfortunately of very limited use in practice as we do not know the variance components, so that we must use the Feasible-GLS estimator.

3.3.1.3 The Feasible-GLS Estimator

The first problem to be solved in estimating the model by Feasible-GLS is to find appropriate estimators for the variance components. The starting point could be the error term u and its decomposition, but because we cannot observe it directly we need to estimate it. These estimates can be based on different consistent estimators of the error components model. The necessary expected values to identify the unknown variances are

$$\begin{aligned} E(u_{it}^2) &= \sigma_\alpha^2 + \sigma_\varepsilon^2 + \sigma_\lambda^2 \\ E((\frac{1}{T}\sum_t u_{it})^2) &= \sigma_\alpha^2 + \frac{1}{T}\sigma_\varepsilon^2 + \frac{1}{T}\sigma_\lambda^2 \\ E((\frac{1}{N}\sum_i u_{it})^2) &= \frac{1}{N}\sigma_\alpha^2 + \frac{1}{N}\sigma_\varepsilon^2 + \sigma_\lambda^2 \end{aligned}$$

so the variance components estimates are

[7] This hypothesis implies that the limits of $\frac{1}{NT}X'B_N X$, $\frac{1}{NT}X'B_T X$ and $\frac{1}{N^2T^2}X'J_{NT}X$ are also finite.

$$\widehat{\sigma}_{\alpha}^2 = \frac{T}{T-1}\left(\frac{\frac{1}{T}\sum_i(\sum_t \widehat{u}_{it})^2}{N-K} - \frac{\sum_i\sum_t \widehat{u}_{it}^2}{NT-K}\right)$$
$$\widehat{\sigma}_{\lambda}^2 = \frac{N}{N-1}\left(\frac{\frac{1}{N}\sum_t(\sum_i \widehat{u}_{it})^2}{T-K} - \frac{\sum_i\sum_t \widehat{u}_{it}^2}{NT-K}\right)$$
$$\widehat{\sigma}_{\varepsilon}^2 = \frac{\sum_i\sum_t \widehat{u}_{it}^2}{NT-K} - \widehat{\sigma}_{\alpha}^2 - \widehat{\sigma}_{\lambda}^2$$

or in another form

$$T\widehat{\sigma}_{\alpha}^2 + \widehat{\sigma}_{\varepsilon}^2 = \frac{\widehat{u}' B_N \widehat{u}}{N-K}$$
$$N\widehat{\sigma}_{\varepsilon}^2 + \widehat{\sigma}_{\varepsilon}^2 = \frac{\widehat{u}' B_T \widehat{u}}{T-K}$$
$$\widehat{\sigma}_{\varepsilon}^2 = \frac{\widehat{u}' W_{NT} \widehat{u}}{(N-1)(T-1)-K-1}$$

where the $\widehat{u}$ residual vector can be obtained by any consistent estimation of the model. The above variance components estimators are consistent under the usual conditions. However, if T is finite ($N \to \infty$) the estimators of σ_{λ}^2 are, of course, inconsistent.

3.3.1.4 The OLS and Within Estimators

When only $N \to \infty$, the OLS estimator is not necessarily consistent, even if we suppose that both

$$\lim_{N\to\infty} \frac{1}{N} X'X = T_{xx}^N$$

and

$$\lim_{N\to\infty} \frac{1}{N} X' W_{NT} X = \widetilde{W}_{xx}^N$$

are finite positive definite matrices (see Trognon (1993)). However, when N and $T \to \infty$ the OLS estimator becomes consistent, but unfortunately, its asymptotic covariance matrix may not be finite if either $\lim_{N\to\infty} \frac{1}{NT} X' B_N X \neq 0$ or $\lim_{N\to\infty} \frac{1}{NT} X' B_T X \neq 0$.

Now let us turn our attention to the Within estimator. It is clear from the decompositions of the error terms that the projection matrix W_{NT} nullifies (similarly as in the case of the one-way model) the individual and time effects. This means that we can get an estimator of model by transforming all the variables with W_{NT} and apply the OLS. We can get this estimator by transforming all variables of the model such as y

$$W_{NT}\, y = \left[y_{it} - \overline{y}_i - \overline{y}_t + \overline{y}\right],$$

and then use the OLS estimator.

As earlier the Within estimator is unbiased, have a normal distribution and a covariance matrix

$$V(\widehat{\beta}_W) = \sigma_\varepsilon^2 (X' W_{NT} X)^{-1}$$

As seen, in the asymptotic case the GLS and the Within estimators are asymptotically equivalent, but in the semi–asymptotic case (N only goes to infinity) the GLS remains more efficient than the within estimator.

3.3.1.5 One-Way Versus Two-Way Error Components

This section investigates the consequences of under-specifying or over-specifying the error components model. Since the one-way and two-way error components models are popular in economics, we focus on the following two cases:

(1) **Under-Specification Case:** In this case the true model is two-way

$$u_{it} = \alpha_i + \lambda_t + \varepsilon_{it} \quad i = 1, \ldots, N; \; t = 1, \ldots, T \tag{3.7}$$

while the estimated model is one-way

$$u_{it} = \alpha_i + \varepsilon_{it} \tag{3.8}$$

with $\alpha_i \sim N(0, \sigma_\alpha^2), \lambda_t \sim N(0, \sigma_\lambda^2)$ and $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$ independent of each other and among themselves. Knowing the true disturbances $u' = (u_{11}, \ldots, u_{1T}, \ldots, u_{N1}, \ldots, u_{NT})$, the Best Quadratic Unbiased (BQU) estimators of the variance components for the one-way model are given by

$$\widehat{\sigma}_\varepsilon^2 = u' W_N \, u / \operatorname{trace}(W_N) \quad \text{and} \quad T\widehat{\sigma}_\alpha^2 + \widehat{\sigma}_\varepsilon^2 = u' B_N \, u / \operatorname{trace}(B_N)$$

Using this fact, one can easily show, (see Baltagi and Li (1991b)) that

$$E(\widehat{\sigma}_\varepsilon^2) = \operatorname{trace}[\Omega W_N / N(T-1)] = \sigma_\varepsilon^2 + \sigma_\lambda^2 , \tag{3.9}$$

which is biased upwards by σ_λ^2. Similarly,

$$E(T\widehat{\sigma}_\alpha^2 + \widehat{\sigma}_\varepsilon^2) = \operatorname{trace}[\Omega B_N / N] = T\sigma_\alpha^2 + \sigma_\varepsilon^2 + \sigma_\lambda^2 , \tag{3.10}$$

which is also biased upwards by σ_λ^2. Substituting $E(\widehat{\sigma}_\varepsilon^2)$ from (3.9) in the left-hand-side of (3.10), one gets $E(\widehat{\sigma}_\alpha^2) = \sigma_\alpha^2$. This shows that knowing the true disturbances, the BQU of σ_ε^2 for the misspecified one-way model is biased upwards, while the BQU of σ_α^2 remains unbiased.

In practice, the true disturbances are not known and may be replaced by the one–way within residuals $\widehat{u}_W = y - X\widehat{\beta}_W$ where $\widehat{\beta}_W$ is the one–way within parameters estimates. In this case

$$\widetilde{\sigma}_\varepsilon^2 = \widehat{u}_W' W_N \widehat{u}_W / N(T-1)$$

and one can easily show that plim $\widetilde{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2 + \sigma_\lambda^2$.

Similarly, substituting $\widehat{u}_W$ for u in $T\widehat{\sigma}_\alpha^2 + \widehat{\sigma}_\varepsilon^2$ we get

$$\widetilde{\sigma}_\alpha^2 = \widehat{u}_W' B_N \widehat{u}_W / NT - \widetilde{\sigma}_\varepsilon^2 / T$$

and one can show that plim $\widetilde{\sigma}_\alpha^2 = \sigma_\alpha^2$. This shows that even if the true disturbances are replaced by the within residuals, the misspecified one-way estimator of σ_ε^2 remains inconsistent, while that of σ_α^2 is consistent.

(2) **Over–Specification Case:** In this case, the true model is one-way, given by (3.8), while the estimated model is two-way, given by (3.7). Knowing the true disturbances, the BQU estimators of the two-way model are given by

$$\widehat{\sigma}_\varepsilon^2 = u' W_{NT} u / (N-1)(T-1) \tag{3.11}$$

$$T\widehat{\sigma}_\alpha^2 + \widehat{\sigma}_\varepsilon^2 = u' B_N u / (N-1) \tag{3.12}$$

$$N\widehat{\sigma}_\lambda^2 + \widehat{\sigma}_\varepsilon^2 = u' B_T u / (T-1) \tag{3.13}$$

(see Amemiya (1971)), where W_{NT}, B_N and B_T are defined as above. Therefore,

$$E(\widehat{\sigma}_\varepsilon^2) = [\Omega W_{NT}] / (N-1)(T-1) = \sigma_\varepsilon^2 \tag{3.14}$$

$$E(T\widehat{\sigma}_\alpha^2 + \widehat{\sigma}_\varepsilon^2) = [\Omega B_N] / (N-1) = T\sigma_\alpha^2 + \sigma_\varepsilon^2 \tag{3.15}$$

$$E(N\widehat{\sigma}_\lambda^2 + \widehat{\sigma}_\varepsilon^2) = [\Omega B_T] / (T-1) = \sigma_\varepsilon^2 . \tag{3.16}$$

Substituting (3.14) in the left-hand-side of (3.15) and (3.16), we get $E(\widehat{\sigma}_\alpha^2) = \sigma_\alpha^2$ and $E(\widehat{\sigma}_\lambda^2) = 0$. This shows that if the true disturbances are known, the BQU estimators of σ_α^2, σ_λ^2 and σ_ε^2 for the misspecified two-way model remain unbiased. If the u_{it}'s are replaced by the two-way within residuals $\widehat{u}_W = y - X\widehat{\beta}_W$ where $\widehat{\beta}_W$ is the two–way within regression estimates given by $\widehat{\beta}_W = (X' W_N X)^{-1} X' W_N y$, then one can show, see Baltagi and Li (1991b), that

$$\begin{aligned} \text{plim}_{N,T\to\infty} \widetilde{\sigma}_\nu^2 &= \text{plim}_{N,T\to\infty} \widehat{u}_W' W_{NT} \widehat{u}_W / (N-1)(T-1) \\ &= \text{plim}_{N,T\to\infty} \text{trace}[\Omega W_{NT} / (N-1)(T-1)] = \sigma_\nu^2 \end{aligned}$$

Similarly, from (3.12) and (3.13), one can show that $\text{plim}_{N,T\to\infty}\ \widetilde{\sigma}_\alpha^2 = \sigma_\alpha^2$ and $\text{plim}_{N,T\to\infty}\ \widetilde{\sigma}_\lambda^2 = 0$.

This shows that if the u_{it}'s are replaced by the two-way within residuals the misspecified two-way variance components estimates remain consistent.

Prucha (1984) showed that as long as the estimator of σ_ε^2 is consistent and the estimators of the other variance components σ_α^2 and σ_λ^2 go to a finite probability limit, as N and T both go to ∞, then the corresponding feasible GLS estimator of β is asymptotically equivalent to the true GLS estimator. This condition is satisfied for the overspecified model but not for the underspecified model.

Deschamps (1991) investigated the consequences of a misspecified error components model on the estimated variances of the regression coefficients. In particular, Deschamps (1991) considered the under-specified case where some error

components are improperly omitted (even though their variances are nonzero), and the remaining variance components are consistently estimated. In this case, Deschamps (1991) shows that the true variances of the estimated regression coefficients are always underestimated in the misspecified model. For the underspecified one–way error components model, with omitted time effects, this inconsistency is unbounded, unless the matrix of regressors satisfies very restrictive assumptions.

3.3.2 Serial Correlation in the Disturbances

The classical error components disturbances assume that the only correlation over time is due to the presence in the panel of the same individual over several periods. This equicorrelation coefficient is given by correl $(u_{it}, u_{is}) = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\varepsilon^2)$ for $t \neq s$. Note that it is the same no matter how far t is from s. This may be a restrictive assumption for economic relationships, like investment or consumption, where an unobserved shock this period will affect the behavioral relationship for at least the next few periods. This type of serial correlation is not allowed for in the simple error components model. Ignoring serial correlation when it is present results in consistent but inefficient estimates of the regression coefficients and biased standard errors. This section introduces serial correlation in the ε_{it}. We illustrate how one can estimate an autoregressive process of order one AR(1), as in the Lillard and Willis (1978) study on earnings.

3.3.2.1 The AR(1) Process

Lillard and Willis (1978) generalized the error components model to the serially correlated case, by assuming that the remainder disturbances (the ε_{it}) follow an AR(1) process. In this case $\alpha_i \sim$ i.i.d.$(0, \sigma_\alpha^2)$, whereas

$$\varepsilon_{it} = \rho \varepsilon_{i,t-1} + \eta_{it} \tag{3.17}$$

$| \rho | < 1$ and $\eta_{it} \sim$ i.i.d.$(0, \sigma_\eta^2)$. The α_i are independent of the ε_{it} and $\varepsilon_{i0} \sim (0, \sigma_\eta^2/(1-\rho^2))$. Baltagi and Li (1991a) derived the corresponding Fuller and Battese (1974) transformation for this model. First, one applies the Prais-Winsten (PW) transformation matrix

$$C = \begin{bmatrix} (1-\rho^2)^{1/2} & 0 & 0 & \cdots & 0 & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & -\rho & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -\rho & 1 \end{bmatrix}$$

to transform the remainder AR(1) disturbances into serially uncorrelated classical errors. For panel data, this has to be applied for N individuals. The transformed regression disturbances are in vector form

$$u^* = (I_N \otimes C)u = (I_N \otimes Ce_T)\alpha + (I_N \otimes C)\varepsilon$$

Using the fact that $Ce_T = (1-\rho)\iota_T^{\varphi\prime}$, where $\iota_T^{\varphi\prime} = (\varphi, e'_{T-1})$ and $\varphi = \sqrt{(1+\rho)/(1-\rho)}$, one can rewrite this as

$$u^* = (1-\rho)(I_N \otimes \iota_T^{\varphi})\alpha + (I_N \otimes C)\varepsilon$$

Therefore, the variance–covariance matrix of the transformed disturbances is

$$\Omega^* = E(u^* u^{*\prime}) = \sigma_\alpha^2 (1-\rho)^2 [I_N \otimes \iota_T^{\varphi} \iota_T^{\varphi\prime}] + \sigma_\eta^2 (I_N \otimes I_T)$$

since $(I_N \otimes C)E(\varepsilon\varepsilon')(I_N \otimes C') = \sigma_\eta^2 (I_N \otimes I_T)$. Alternatively, this can be rewritten as

$$\Omega^* = d^2 \sigma_\alpha^2 (1-\rho)^2 [I_N \otimes \iota_T^{\varphi} \iota_T^{\varphi\prime}/d^2] + \sigma_\eta^2 (I_N \otimes I_T)$$

where $d^2 = \iota_T^{\varphi\prime} \iota_T^{\varphi} = \varphi^2 + (T-1)$ or equivalently,

$$\Omega^* = \sigma_\kappa^2 (I_N \otimes \iota_T^{\varphi} \iota_T^{\varphi\prime}/d^2) + \sigma_\eta^2 (I_N \otimes (I_T - \iota_T^{\varphi} \iota_T^{\varphi\prime}/d^2))$$

where $\sigma_\kappa^2 = d^2 \sigma_\alpha^2 (1-\rho)^2 + \sigma_\eta^2$. Therefore

$$\begin{aligned} \sigma_\eta \Omega^{*-1/2} &= (\sigma_\eta/\sigma_\kappa)(I_N \otimes \iota_T^{\varphi} \iota_T^{\varphi\prime}/d^2) + (I_N \otimes (I_T - \iota_T^{\varphi} \iota_T^{\varphi\prime}/d^2)) \\ &= I_N \otimes I_T - \theta_\kappa (I_N \otimes \iota_T^{\varphi} \iota_T^{\varphi\prime}/d^2) \end{aligned}$$

where $\theta_\kappa = 1 - (\sigma_\eta/\sigma_\kappa)$.

Premultiplying the PW transformed observations $y^* = (I_N \otimes C)y$ by $\sigma_\eta \Omega^{*-1/2}$ one gets $y^{**} = \sigma_\eta \Omega^{*-1/2} y^*$. The typical elements of $y^{**} = \sigma_\eta \Omega^{*-1/2} y^*$ are given by

$$(y_{i1}^* - \theta_\kappa \varphi \beta_i, y_{i2}^* - \theta_\kappa \beta_i, \ldots, y_{iT}^* - \theta_\kappa \beta_i)'$$

where $\beta_i = [(\varphi y_{i1}^* + \sum_2^T y_{it}^*)/d^2]$ for $i = 1, \ldots, N$. The first observation gets special attention in the AR(1) error components model. First, the PW transformation gives it a special weight $\sqrt{1-\rho^2}$ in y^*. Second, the Fuller and Battese transformation also gives it a special weight $\varphi = \sqrt{(1+\rho)/(1-\rho)}$ in computing the weighted average β_i and the pseudo-difference. Note that

(i) if $\rho = 0$, then $\varphi = 1$, $d^2 = T$, $\sigma_\kappa^2 = \sigma_1^2$ and $\theta_\kappa = \theta$. Therefore, the typical element of y_{it}^{**} reverts to the familiar $(y_{it} - \theta \bar{y}_{i.})$ transformation for the one-way error component model with no serial correlation.

(ii) If $\sigma_\alpha^2 = 0$, then $\sigma_\kappa^2 = \sigma_\eta^2$ and $\theta_\kappa = 0$. Therefore, the typical element of y_{it}^{**} reverts to the PW transformation y_{it}^*.

The Best Quadratic Unbiased (BQU) estimators of the variance components are given by

$$\widehat{\sigma}_{\eta}^2 = u^{*\prime}(I_N \otimes (I_N \otimes \iota_T^{\varphi}\iota_T^{\varphi\prime}/d^2)u^*/N(T-1)$$
$$\text{and} \quad \widehat{\sigma}_{\kappa}^2 = u^{*\prime}(I_N \otimes \iota_T^{\varphi}\iota_T^{\varphi\prime}/d^2)u^*/N$$

of σ_{η}^2 and σ_{κ}^2 respectively.

Baltagi and Li (1991a) suggest estimating ρ from Within residuals $\widetilde{\varepsilon}_{it}$ as $\widetilde{\rho} = \sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{\varepsilon}_{i,t}\widetilde{\varepsilon}_{i,t-1}/\sum_{i=1}^{N}\sum_{t=2}^{T}\widetilde{\varepsilon}_{i,t-1}^2$. Then, $\widehat{\sigma}_{\eta}^2$ and $\widehat{\sigma}_{\kappa}^2$ are estimated by substituting OLS residuals $\hat{u}^*$ from the PW transformed equation using $\widetilde{\rho}$. Using Monte Carlo experiments, Baltagi and Li (1997) found that $\widetilde{\rho}$ performs poorly for small T and recommended an alternative estimator of ρ which is based on the autocovariance function $Q_s = E(u_{it}u_{i,t-s})$. For the AR(1) model, it is easy to show that $Q_s = \sigma_{\alpha}^2 + \sigma_{\eta}^2\rho^s$. From Q_0, Q_1 and Q_2, one can easily show that $\rho + 1 = (Q_0 - Q_2)/(Q_0 - Q_1)$. Hence, a consistent estimator of ρ (for large N) is given by

$$\widehat{\rho} = \frac{\widetilde{Q}_0 - \widetilde{Q}_2}{\widetilde{Q}_0 - \widetilde{Q}_1} - 1 = \frac{\widetilde{Q}_1 - \widetilde{Q}_2}{\widetilde{Q}_0 - \widetilde{Q}_1}$$

where $\widetilde{Q}_s = \sum_{i=1}^{N}\sum_{t=s+1}^{T}\widehat{u}_{it}\widehat{u}_{i,t-s}/N(T-s)$ and $\widehat{u}_{it}$ denotes the OLS residuals. $\widehat{\sigma}_{\eta}^2$ and $\widehat{\sigma}_{\kappa}^2$ are estimated by substituting OLS residuals $\hat{u}^*$ from the PW transformed equation using $\widehat{\rho}$ rather than $\widetilde{\rho}$.

Therefore, the estimation of an AR(1) serially correlated error components model is considerably simplified by

(i) applying the PW transformation in the first step, as is usually done in the time-series literature, and
(ii) subtracting a pseudo-average from these transformed data in the second step.

3.3.2.2 Kmenta's Approach

(1) The Common ρ Case

In this case the disturbances are assumed to follow a first order autoregressive process with the same ρ but different variances for different cross sections, i.e.,

$$u_{it} = \rho u_{i,t-1} + \eta_{it} \tag{3.18}$$

with $\eta_{it} \sim N(0, \sigma_{\eta i}^2)$ and $u_{i0} \sim N(0, \sigma_{\eta i}^2/(1-\rho^2))$. The estimation method proposed is to correct for serial correlation in the first step and heteroskedasticity in the second step. This is accomplished by estimating ρ by $\widehat{\rho} = \Sigma\Sigma\widehat{u}_{it}\widehat{u}_{i,t-1}/\Sigma\Sigma\widehat{u}_{i,t-1}^2$ with $\widehat{u}_{it}$ denoting the OLS residuals on the pooled model. Next, the Prais–Winsten transformation is applied, i.e., $y_{it}^* = y_{it} - \widehat{\rho}y_{i,t-1}$ for $t = 2, \ldots, T$, and $y_{i,1}^* = (1 - \widehat{\rho}^2)^{1/2}y_{i,1}$ with a similar transformation on the X_{it}'s. y_{it}^* is regressed on the X_{it}^*'s and the residuals $\widehat{u}_{it}^*$'s are formed. Estimates of the variances are obtained as

$\widehat{\sigma}^2_{\eta i} = \Sigma \widehat{u}^{*2}_{it}/(T-K)$ for $i = 1, 2, \ldots, N$, and $y^{**}_{it} = y^{*}_{it}/\widehat{\sigma}_{\eta i}$ and X^{**}_{it}'s are formed. Finally y^{**}_{it} is regressed on the X^{**}_{it}'s. This procedure when iterated until convergence will lead to maximum likelihood estimates.

(2) The Varying ρ Case

Kmenta (1986) also suggested to consider cases where the serial correlation can vary across individuals

$$u_{it} = \rho_i u_{i,t-1} + \eta_{it} \tag{3.19}$$

with η_{it} defined above.

Maintaining the assumption of independence across individuals, the N ρ_i's are then estimated by $\widehat{\rho}_i = \Sigma \widehat{u}_{it}\widehat{u}_{i,t-1}/\Sigma \widehat{u}^2_{i,t-1}$ for $i = 1, 2, \ldots, N$, and the remaining steps are the same as above.

Kmenta (1986) also considered the situation where some correlation may exist between individuals. In this case

$$E(u_{it}u_{jt}) = \sigma_{ij} \quad \text{for} \quad i, j = 1, 2, \ldots, N,$$

and $E(\varepsilon_{it}\varepsilon_{jt}) = \phi_{ij}$ with $\sigma_{ij} = \phi_{ij}/(1-\rho_i\rho_j)$. The variance–covariance matrix is now $\Omega = [\sigma_{ij}V_{ij}]$ where

$$V_{ij} = \begin{bmatrix} 1 & \rho_j & \rho_j^2 & \cdots & \rho_j^{T-1} \\ \rho_i & 1 & \rho_j & & \vdots \\ \rho_i^2 & & 1 & & \\ \vdots & & & \ddots & \rho_j \\ \rho_i^{T-1} & \cdots & & \rho_i & 1 \end{bmatrix}.$$

Estimates of the ρ_i's and ϕ_{ij}'s are obtained as before with $\widehat{\phi}_{ij} = \Sigma e^*_{it}e^*_{jt}/(T-K)$. Rather than applying GLS which inverts and $NT \times NT$ matrix, Kmenta (1986) suggests running GLS on the transformed model, i.e., using $(\underline{y}^*, X^*)$ as follows:

$$\widehat{\beta} = (X^{*\prime}\widehat{\Phi}^{-1}X^*)^{-1}(X^{*\prime}\widehat{\Phi}^{-1}\underline{y}^*)$$

where $\widehat{\Phi} = \widehat{\phi} \otimes I_T$, $\widehat{\phi} = [\widehat{\phi}_{ij}]$. $\widehat{\phi}$ is $N \times N$ and if N is larger than T, which is the usual case in economics, this $\widehat{\phi}$ is singular. Also, $\widehat{\beta}$ is not GLS since the Prais–Winsten transformation does not give the right first element of V_{ij} for $i \neq j$. Kmenta suggests ignoring the first observation, i.e., applying Cochrane–Orcutt. The transformation would be correct in this case but we lose N observations, one for each cross section. This could be a lot of observations lost for panels with large N.

3.3.3 Two-Way Error Components vs Kmenta's Approach

The usual error components model as well as the Kmenta technique allow for serial correlation, but in the usual error components model this serial correlation is

constant across time, whereas it decays over time with the Kmenta technique. In its most general case, the Kmenta technique allows for correlation among the different cross sections, whereas the error components technique assumes a lot of independence among the α's, λ's and ε's. Moreover, the usual error components model has homoskedastic disturbances, whereas the Kmenta technique has heteroskedastic ones. Also, the Kmenta technique estimates a lot of auxiliary parameters, for e.g., N σ's and one ρ in the case of the common rho method, N σ's and N ρ's in case of the varying rho method, and $N(N+1)/2$ σ's and N ρ's in the cross-sectionally correlated time-wise autoregressive case. In the fixed effects model, one estimates $(N-1)$ α's and $(T-1)$ ε's, but in the two-way random effects model, one only estimates three variances.

The advantages of both methods are the gains from pooling a larger data set and more variation to explain the underlying economic relationship. However, as usual in economics, the true structure of the disturbances is not known, and the researcher may be at a disadvantage if the wrong error structure is chosen. Given this background, Baltagi (1986) posed the following basic question: Under the best possible situation of applying one technique of pooling, how does the other technique perform? This question is relevant given the wide use of the Kmenta and the error components techniques and their easy accessibility on computer. The performance of these methods is compared by means of Monte-Carlo experiments. First, data are generated with both serially correlated and cross-sectionally heteroskedastic disturbances and both the Kmenta and the familiar variance components methods are applied. Next, data are generated with error components disturbances, and again both techniques of estimation are applied. For $N = 25$ and $T = 10$, Baltagi (1986) shows that the error components procedure is more robust to this kind of misspecification than the Kmenta technique. This result should be tempered by the fact that N is large and T is small. This means, that, in the Kmenta case, one is estimating a lot of a auxiliary parameters with a short time series. If T is large and N is small, the Kmenta technique is expected to perform better. Most panels in economics, however, are of the type where N is much larger than T. Also, the Monte-Carlo results show that the error components Feasible GLS estimators differ from each other when the model is misspecified. Negative estimates of the variance components and non–stationary ρ's occur when the model is misspecified. Finally, OLS performs better than the wrong Feasible GLS estimator, but worse than the correct Feasible GLS estimator. Note that a robust variance–covariance matrix for the OLS estimator under the Kmenta model was proposed by Beck and Katz (1995).

3.3.4 Heteroskedasticity in the Disturbances

The standard error components model assumes that the regression disturbances are homoskedastic with the same variance across time and individuals. This may be a restrictive assumption for panels, where the cross-sectional units may be of varying

size and as a result may exhibit different variation. For example, when dealing with different size countries or firms, one should expect to find heteroskedasticity in the disturbance terms. Assuming homoskedastic disturbances when heteroskedasticity is present will still result in consistent estimates of the regression coefficients, but these estimates will not be efficient. Also, the standard errors of these estimates will be biased and one should compute robust standard errors correcting for the possible presence of heteroskedasticity. In this section, we relax the assumption of homoskedasticity of the disturbances and introduce heteroskedasticity through the α_i as first suggested by Mazodier and Trognon (1978). Next, we suggest an alternative heteroskedastic error components specification, where only the ε_{it} are heteroskedastic. We derive the true GLS transformation for these two models. We also consider two adaptive heteroskedastic estimators based on these models where the heteroskedasticity is of unknown form. These adaptive heteroskedastic estimators were suggested by Li and Stengos (1994) and Roy (2002).

Mazodier and Trognon (1978) generalized the homoskedastic error components model to the case where the α_i are heteroskedastic, i.e., $\alpha_i \sim (0, \sigma^2_{\alpha i})$ for $i = 1, \ldots, N$, but $\varepsilon_{it} \sim$ i.i.d.$(0, \sigma^2_\varepsilon)$. In vector form, $\alpha \sim (0, \Sigma_\alpha)$ where $\Sigma_\alpha = \text{diag}[\sigma^2_{\alpha i}]$ is a diagonal matrix of dimension $N \times N$, and $\varepsilon \sim (0, \sigma^2_\varepsilon I_{NT})$..... Therefore, the resulting variance–covariance of the disturbances is given by

$$\Omega = \text{diag}[\sigma^2_{\alpha i}] \otimes J_T + \text{diag}[\sigma^2_\varepsilon] \otimes I_T$$

where $\text{diag}[\sigma^2_\varepsilon]$ is also of dimension $N \times N$. This can be rewritten as follows

$$\Omega = \text{diag}[\tau_i^2] \otimes \frac{J_T}{T} + \text{diag}[\sigma^2_\varepsilon] \otimes (I_T - \frac{J_T}{T})$$

with $\tau_i^2 = T\sigma^2_{\alpha i} + \sigma^2_\varepsilon$. In this case,

$$\sigma_\varepsilon \Omega^{-1/2} = (\text{diag}[\sigma_\varepsilon / \tau_i] \otimes \frac{J_T}{T}) + W_N$$

Hence, $y^* = \sigma_\nu \Omega^{-1/2} y$ has a typical element $y^*_{it} = y_{it} - \theta_i \bar{y}_i$ where $\theta_i = 1 - (\sigma_\varepsilon / \tau_i)$ for $i = 1, \ldots, N$.

Baltagi and Griffin (1988) provided Feasible GLS estimators including Rao's (1970, 1972) MINQUE estimators for this model. Phillips (2003) argues that this model suffers from the incidental parameters problem and the variance estimates of α_i (the $\sigma^2_{\alpha i}$) cannot be estimated consistently, so there is no guarantee that Feasible GLS and true GLS will have the same asymptotic distributions. Instead, he suggests a stratified error components model where the variances change across strata and provides an EM algorithm to estimate it. It is important to note that Mazodier and Trognon (1978) had already suggested stratification in a two-way heteroskedastic error component model. Also, that one can specify parametric variance functions which avoid the incidental parameter problem and then apply the GLS transformation described above. As in the cross-section heteroskedastic case, one has to know the variables that determine heteroskedasticity, but not necessarily the form.

Adaptive estimation of heteroskedasticity of unknown form has been suggested for this model by Roy (2002). This follows similar literature on adaptive estimation for cross-section data.

Alternatively, one could keep the α_i homoskedastic with $\alpha_i \sim$ i.i.d.$(0, \sigma_\alpha^2)$ and impose the heteroskedasticity on the ε_{it}, i.e., $\varepsilon_{it} \sim (0, \sigma_{\varepsilon i}^2)$ (see problem 88.2.2 by Baltagi (1988) and its solution by Wansbeek (1989)). In this case, one obtains

$$\Omega = E(uu') = \text{diag}[\sigma_\alpha^2] \otimes J_T + \text{diag}[\sigma_{\varepsilon i}^2] \otimes I_T$$

which can be rewritten as

$$\Omega = \text{diag}[T\sigma_\alpha^2 + \sigma_{\varepsilon i}^2] \otimes \frac{J_T}{T} + \text{diag}[\sigma_{\varepsilon i}^2] \otimes (I_T - \frac{J_T}{T})$$

and

$$\Omega^{-1/2} = \text{diag}[1/\tau_i] \otimes \frac{J_T}{T} + \text{diag}[1/\sigma_{\varepsilon i}] \otimes (I_T - \frac{J_T}{T})$$

and $y^* = \Omega^{-1/2} y$ has a typical element

$$y_{it}^* = (\bar{y}_i/\tau_i) + (y_{it} - \bar{y}_i)/\sigma_{\varepsilon i} \, .$$

Upon rearranging terms, we get

$$y_{it}^* = \frac{1}{\sigma_{\varepsilon i}}(y_{it} - \theta_i \bar{y}_i) \quad \text{where} \quad \theta_i = 1 - (\sigma_{\varepsilon i}/\tau_i)$$

Estimators for this one-way random effects model with unequal error variances and no regressors has been studied extensively in the statistics literature, see Rao, Kaplan and Cochran (1981) for a good review.

One can argue that heteroskedasticity will contaminate both α_i and ε_{it} and it is hard to claim that it is in one component and not the other. Randolph (1988) gives the GLS transformation for a more general heteroskedastic model where *both* the α_i and the ε_{it} are assumed heteroskedastic in the context of an unbalanced panel. In this case, the var$(\alpha_i) = \sigma_{\alpha i}^2$ and $E(\varepsilon\varepsilon') = \text{diag}[\sigma_{it}^2]$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T_i$. More recently, Li and Stengos (1994) considered the case where $\alpha_i \sim$ i.i.d. $\left(0, \sigma_\alpha^2\right)$ and $E\left[\varepsilon_{it} | x_{it}'\right] = 0$ with Var $\left[\varepsilon_{it} | x_{it}'\right] = \gamma(x_{it}') \equiv \gamma_{it}$. So that the heteroskedasticity is on the remainder error term and it is of an unknown form.

Therefore $\sigma_{it}^2 = E\left[u_{it}^2 | x_{it}'\right] = \sigma_\alpha^2 + \gamma_{it}$ and the proposed estimator of σ_α^2 is given by

$$\widehat{\sigma}_\alpha^2 = \frac{\sum_{i=1}^{N} \sum_{t \neq s}^{T} \widehat{u}_{it} \widehat{u}_{is}}{NT\,(T-1)}$$

where $\widehat{u}_{it}$ denotes the OLS residual. Also,

$$\widehat{\gamma}_{it} = \frac{\sum_{j=1}^{N}\sum_{s=1}^{T} \widehat{u}_{js}^2 K_{it,js}}{\sum_{j=1}^{N}\sum_{s=1}^{T} K_{it,js}} - \widehat{\sigma}_{\alpha}^2$$

where the kernel function is given by $K_{it,js} = K\left(\frac{x_{it}' - x_{js}'}{h}\right)$ and h is the smoothing parameter. These estimators of the variance components are used to construct a feasible adaptive GLS estimator of β which they denote by GLSAD. The computation of their Feasible GLS estimator is simplified into an OLS regression using a recursive transformation that reduces the general heteroskedastic error components structure into classical errors, see Li and Stengos (1994) for details.

Roy (2002) considered the alternative heteroskedastic model $E[\alpha_i | \bar{x}_i'] = 0$ with

$$\text{Var}\left[\alpha_i | \bar{x}_i'\right] = \omega\left(\bar{x}_i'\right) \equiv \omega_i$$

with $\bar{x}_i' = \sum_{t=1}^{T} x_{it}'/T$ and $v_{it} \sim \text{i.i.d.}\left(0, \sigma_v^2\right)$. So that the heteroskedasticity is on the individual specific error component and it is of an unknown form. Roy (2002) used the usual estimator of σ_v^2 which is the MSE of the Within regression, and this can be written as

$$\widehat{\sigma}_{\varepsilon}^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} \left[\left(y_{it} - \bar{y}_{i.}\right) - \left(x_{it}' - \bar{x}_i'\right)\beta_W\right]^2}{N(T-1) - k}$$

where β_W is the fixed effects or within estimator of β. Also

$$\widehat{\omega}_i = \frac{\sum_{j=1}^{N}\sum_{t=1}^{T} \widehat{u}_{jt}^2 K_{i.,j.}}{\sum_{j=1}^{N}\sum_{t=1}^{T} K_{i.,j.}} - \widehat{\sigma}_{\varepsilon}^2$$

where the kernel function is given by

$$K_{i.,j.} = K\left(\frac{\bar{x}_i' - \bar{x}_j'}{h}\right)$$

Using these estimators of the variance components, Roy (2002) computed a Feasible GLS estimator using the transformation derived by Baltagi and Griffin (1988). This was denoted by EGLS.

Both Li and Stengos (1994) and Roy (2002) performed Monte Carlo experiments based on the simple regression model with one regressor. They compared the following estimators:

(1) OLS;
(2) Fixed effects or within estimator (Within);
(3) the conventional GLS estimator for the one way error components model that assumes the error term u_{it} is homoskedastic (GLSH); and
(4) their own adaptive heteroskedastic estimator denoted by (EGLS) for Roy (2002) and (GLSAD) for Li and Stengos (1994).

Li and Stengos (1994) found that their adaptive estimator outperforms all the other estimators in terms of relative MSE with respect to true GLS for $N = 50, 100$ and $T = 3$ and for moderate to severe degrees of heteroskedasticity. Roy (2002) also found that her adaptive estimator performs well, although it was outperformed by fixed effects in some cases where there were moderate and severe degrees of heteroskedasticity. Recently, Baltagi, Bresson and Pirotte (2005) checked the sensitivity of the two proposed adaptive heteroskedastic estimators under misspecification of the form of heteroskedasticity. In particular, they ran Monte Carlo experiments using the heteroskedasticity set up of Li and Stengos (1994) to see how the misspecified Roy (2002) estimator performs. Next, they used the heteroskedasticity set up of Roy (2002) to see how the misspecified Li and Stengos (1994) estimator performs. They also checked the sensitivity of these results to the choice of the smoothing parameters, the sample size and the degree of heteroskedasticity. Baltagi, Bresson and Pirotte (2005) found that in terms of loss in efficiency, misspecifying the adaptive form of heteroskedasticity can be costly when the Li and Stengos (1994) model is correct and the researcher performs the Roy (2002) estimator. This loss in efficiency is smaller when the true model is that of Roy (2002) and one performs the Li and Stengos (1994) estimator. The latter statement is true as long as the choice of bandwidth is not too small. Both papers also reported the 5% size performance of the estimated t-ratios of the slope coefficient. Li and Stengos (1994) found that only GLSAD had the correct size while OLS, GLSH and Within over-rejected the null hypothesis. Roy (2002) found that GLSH and EGLS had the correct size no matter what choice of h was used. Baltagi, Bresson and Pirotte (2005) found that OLS and GLSAD (small h) tend to over-reject the null when true no matter what form of adaptive heteroskedasticity. In contrast, GLSH, EGLS and Within have size not significantly different from 5% when the true model is that of Roy (2002) and slightly over-reject (7–8%) when the true model is that of Li and Stengos (1994).

3.4 Testing

As in any other econometric context, the properties of the estimators considered in this chapter rely on the validity of the assumptions made. It is then essential to check whether these assumptions can be considered as validated, or if they must be rejected. A first question to answer is whether there exist individual effects at all or not.

3.4.1 Testing for the Absence of Individual Effects

Several tests have been proposed in the literature in order to check for the absence of individual (random) effects. We shall limit ourselves to three of them, which present the advantage of being quite simple to implement, and which nevertheless have rather good properties.

3.4.1.1 The Analysis of Variance / Fisher's Test

A first way of checking for the absence of individual effects consists of testing for the nullity of their variance σ_α^2:

$$\begin{aligned} H_0 : \quad & \sigma_\alpha^2 = 0 \\ \text{against } H_1 : \quad & \sigma_\alpha^2 \neq 0. \end{aligned}$$

This test is very easy to implement as long as we have run the Within and Between regressions from which we get the estimated residual variances $\widehat{\sigma}_w^2$ $(= \widehat{\sigma}_\varepsilon^2)$ and $\widehat{\sigma}_b^2 = \widehat{\sigma}_\alpha^2 + \widehat{\sigma}_\varepsilon^2/T$. Under the normality assumption,

$$(N(T-1) - k_w)\frac{\widehat{\sigma}_w^2}{\sigma_\varepsilon^2}$$

is distributed as a χ^2 with $(N(T-1) - k_w)$ degrees of freedom, and

$$(N - k_b)\frac{\widehat{\sigma}_b^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2/T} = (N - k_b)\frac{T\widehat{\sigma}_b^2}{T\sigma_\alpha^2 + \sigma_\varepsilon^2}$$

is also distributed as a χ^2 but with $(N - k_b)$ degrees of freedom. As a consequence,

$$\frac{\sigma_\varepsilon^2}{T\sigma_\alpha^2 + \sigma_\varepsilon^2}\frac{T\widehat{\sigma}_b^2}{\widehat{\sigma}_w^2} \quad \rightarrow F(N - k_b\ ,\ N(T-1) - k_w)\ .$$

Then, under the null hypothesis, $H_0 :\ \sigma_\alpha^2 = 0$, we have:

$$\frac{T\widehat{\sigma}_b^2}{\sigma_\varepsilon^2} \quad \rightarrow F(N - k_b\ ,\ N(T-1) - k_w)\ .$$

Consequently, one will reject H_0 when this statistics is larger than the fractile of the Fisher distribution with $(N - k_b\ ,\ N(T-1) - k_w)$ degrees of freedom. Simply stated, if T times the Between regression individual variance (computed over N observations) is larger than the residual variance of the Within regression, one must reject the absence of individual effects: such effects do exist.

3.4.1.2 The Lagrange Multiplier Test

Breusch and Pagan (1979) have proposed to use the Lagrange multiplier test to check for the absence of individual effects. Their idea is that, when there are no such effects, the disturbances of the model are completely idiosyncratic. In that situation, the variance of the disturbances should not significantly differ from that of their individual means ($\overline{\varepsilon}_i$), once the necessary correction of the scale effect for the variance of a mean has been made. Then, under such an assumption, the statistics

$$g = \frac{NT}{2(T-1)} \left[\frac{\sum_{i=1}^{N} (T\overline{\widehat{\varepsilon}}_i)^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\varepsilon}_{it}^2} - 1 \right]^2$$

is asymptotically distributed as a χ^2 with 1 degree of freedom. Consequently, if this statistics, computed from the OLS regression residuals is greater than 3.84 (when testing at 5%), one will reject the null of absence of individual specific effects. In the opposite case, one should accept this assumption.

3.4.1.3 Honda's Test

A drawback of the above test as proposed by Breusch and Pagan (1979) is that this is a two-sided test, while a variance should be either null or positive. In order to circumvent this problem, Honda (1985) has suggested a very simple one-sided test that just amounts to consider the square root of the Breusch–Pagan statistics. He showed that, under the null of absence of individual effects, this square root is distributed as a normal. Then, one should reject the null hypothesis as soon as the statistics

$$g = \sqrt{\frac{NT}{2(T-1)}} \left[\frac{\sum_{i=1}^{N} (T\overline{\widehat{\varepsilon}}_i)^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\varepsilon}_{it}{}^2} - 1 \right]$$

is greater than 1.64. This test then leads to reject the null a bit more often than the Breusch–Pagan test would do.

3.4.2 Testing for Uncorrelated Effects: Hausman's Test

One of the most disputable assumptions underlying the error components model is the absence of correlation between the regressors and the individual effects. Indeed, in many circumstances, this is a quite untenable assumption. It is then important to check for the validity of this assumption as such a correlation would lead to the inconsistency of most of the estimators of the error components model,[8] with the notable exception of the Within estimator. Since the latter is based on a transformation

[8] The OLS, Feasible-GLS and Between estimators are biased and inconsistent when only $N \to \infty$. Feasible-GLS remain consistent when both N and $T \to \infty$.

that discards the individual effects from the model, this makes the assumption of their uncorrelation with the regressors irrelevant to the unbiasedness and consistency of this estimator.

Hausman (1978) has suggested a test that exploits the fact that a couple of estimators may be defined in such a way that one ($\widehat{\beta}(1)$) is consistent both under H_0 and H_1 while the other one ($\widehat{\beta}(2)$) is consistent and efficient only if H_0 is true and inconsistent otherwise. Then, getting close estimates $\widehat{\beta}(1)$ and $\widehat{\beta}(2)$ is an indication that H_0 is true while getting very different estimates $\widehat{\beta}(1)$ and $\widehat{\beta}(2)$ must be seen as an indication that H_0 is not validated. More specifically, Hausman has shown that under H_0.

$$Q_H = (\widehat{\beta}(1) - \widehat{\beta}(2))'[\widehat{V(\widehat{\beta}(1))} - \widehat{V(\widehat{\beta}(2))}]^{-1}(\widehat{\beta}(1) - \widehat{\beta}(2))$$

is asymptotically distributed (when $N \to \infty$) as a Chi-Squared with $\dim(\beta)$ degrees of freedom. If Q_H is larger than the fractile of the $\chi^2_{(\dim(\beta))}$ distribution, one must reject H_0; while this assumption is accepted otherwise.

In our current context, we can choose the Within estimator as $\widehat{\beta}(1)$ while the Feasible-GLS estimator is the choice to be made for $\widehat{\beta}(2)$. Then, the statistics to be computed is given by

$$Q_H = (\widehat{\beta}_w - \widehat{\beta}_{\text{fgls}})'[\widehat{V(\widehat{\beta}_w)} - \widehat{V(\widehat{\beta}_{\text{fgls}})}]^{-1}(\widehat{\beta}_w - \widehat{\beta}_{\text{fgls}}) .$$

If Q_H is greater than the fractile of a $\chi^2_{(k_w)}$ where k_w is the number of regressors in the Within regression, one should reject H_0: the absence of correlation between the regressors and the individual effects must be rejected. As a consequence, while the Within estimator is consistent, the Feasible-GLS nor the other estimators (OLS and Between) are consistent in this case.

It is worthwhile noticing that this test can also be conducted in alternative ways. Indeed, Hausman and Taylor (1981) have shown that one can answer the same question by comparing

$$Q_{\text{fgls},b} = (\widehat{\beta}_b - \widehat{\beta}_{\text{fgls}})'[\widehat{V(\widehat{\beta}_b)} - \widehat{V(\widehat{\beta}_{\text{fgls}})}]^{-1}(\widehat{\beta}_b - \widehat{\beta}_{\text{fgls}})$$

or, alternatively

$$Q_{w,b} = (\widehat{\beta}_b - \widehat{\beta}_w)'[\widehat{V(\widehat{\beta}_b)} + \widehat{V(\widehat{\beta}_w)}]^{-1}(\widehat{\beta}_b - \widehat{\beta}_w) .$$

to a $\chi^2_{(k_w)}$. Indeed, these three statistics are (Q_s $N \to \infty$) numerically identical.

3.4.3 Testing for Serial Correlation

In this section, we address the problem of *jointly* testing for serial correlation and individual effects. Baltagi and Li (1995) derived three LM statistics for an error components model with first-order serially correlated errors. The first LM statistic

jointly tests for zero first-order serial correlation and random individual effects. The second LM statistic tests for zero first-order serial correlation assuming fixed individual effects, and the third LM statistic tests for zero first-order serial correlation assuming random individual effects. In all three cases, Baltagi and Li (1995) showed that the corresponding LM statistic is the *same* whether the alternative is AR(1) or MA(1).

Let us assume the disturbances to follow a one-way error components model where $\alpha_i \sim$ i.i.d.$(0, \sigma_\alpha^2)$ and the remainder disturbance follows a stationary AR(1) process: $\varepsilon_{it} = \rho\varepsilon_{i,t-1} + \eta_{it}$ with $|\rho| < 1$, or an MA(1) process: $\varepsilon_{it} = \varepsilon_{it} + \lambda\varepsilon_{i,t-1}$ with $|\lambda| < 1$, and $\varepsilon_{it} \sim$ i.i.d.$(0, \sigma_\varepsilon^2)$. The joint LM test statistic for H_1^a: $\sigma_\alpha^2 = 0$; $\lambda = 0$ is the same as that for H_1^b: $\sigma_\alpha^2 = 0$; $\rho = 0$ and is given by

$$LM_1 = \frac{NT^2}{2(T-1)(T-2)}[A^2 - 4AB + 2TB^2]$$

where $\widehat{u}$ denote OLS residuals, $A = [\widehat{u}'(I_N \otimes J_T)\widehat{u}/(\widehat{u}'\widehat{u})] - 1$ and $B = (\widehat{u}'\widehat{u}_{-1}/\widehat{u}'\widehat{u})$. This is asymptotically distributed (for large N) as χ_2^2 under H_1^a.

Note that the A^2 term is the basis for the LM test statistic for H_2: $\sigma_\alpha^2 = 0$ assuming there is no serial correlation (see Breusch and Pagan, 1980 or Sect. 4.1.2). In fact, $\text{LM}_2 = \sqrt{NT/2(T-1)}A$ is asymptotically distributed (for large N) as $N(0,1)$ under H_2 against the one-sided alternative H_2'; $\sigma_\alpha^2 > 0$. Also, the B^2 term is the basis for the LM test statistic for H_3: $\rho = 0$ (or $\lambda = 0$) assuming there are no individual effects (see Breusch and Godfrey, 1981). In fact, $\text{LM}_3 = \sqrt{NT^2/(T-1)}B$ is asymptotically distributed (for large N) as $N(0,1)$ under H_3 against the one-sided alternative H_3'; ρ (or λ) > 0. The presence of an interaction term in the joint LM test statistic, emphasizes the importance of the joint test when both serial correlation and random individual effects are suspected. However, when T is large the interaction term becomes negligible.

Also, Baltagi and Li (1995) derived two extensions of the Burke, Godfrey and Termayne (1990) AR(1) vs MA(1) test from the time-series to the panel data literature. The first extension tests the null of AR(1) disturbances against MA(1) disturbances, and the second the null of MA(1) disturbances against AR(1) disturbances in an error components model. These tests are computationally simple requiring only OLS or Within residuals.

3.4.4 Testing for Heteroskedasticity

Verbon (1980) derived a Lagrange multiplier test for the null hypothesis of homoskedasticity against the heteroskedastic alternative $\alpha_i \sim \left(0, \sigma_{\alpha_i}^2\right)$ and $\varepsilon_{it} \sim \left(0, \sigma_{\varepsilon_i}^2\right)$. In Verbon's model, however, $\sigma_{\alpha_i}^2$ and $\sigma_{\varepsilon_i}^2$ are, up to a multiplicative constant, *identical* parametric functions of time invariant exogenous variables Z_i, i.e., $\sigma_{\alpha_i}^2 = \sigma_\alpha^2 f(Z_i\theta_2)$ and $\sigma_{\varepsilon_i}^2 = \sigma_v^2 f(Z_i\theta_1)$. Lejeune (1996) on the other hand, dealt with maximum likelihood estimation and Lagrange multiplier testing of a

general heteroskedastic one-way error components regression model assuming that $\alpha_i \sim \left(0, \sigma^2_{\alpha_i}\right)$ and $\varepsilon_{it} \sim \left(0, \sigma^2_{\varepsilon_{it}}\right)$ where $\sigma^2_{\alpha_i}$ and $\sigma^2_{\varepsilon_i}$ are *distinct* parametric functions of exogenous variables Z_{it} and F_i, i.e., $\sigma^2_{\varepsilon_i} = \sigma^2_{\varepsilon} h_{\varepsilon}\left(Z_{it}\theta_1\right)$ and $\sigma^2_{\alpha_i} = \sigma^2_{\alpha} h_{\alpha}\left(F_i\theta_2\right)$. In the context of incomplete panels, Lejeune (1996) derived two joint LM tests for no individual effects and homoskedasticity in the remainder error term. The first LM test considers a random effects one-way error components model with $\alpha_i \sim$ i.i.d. $\left(0, \sigma^2_{\alpha}\right)$ and a remainder error term that is heteroskedastic $\varepsilon_{it} \sim N\left(0, \sigma^2_{\varepsilon_{it}}\right)$ with $\sigma^2_{\varepsilon_{it}} = \sigma^2_{\varepsilon} h_{\varepsilon}\left(Z_{it}\theta_1\right)$. The joint hypothesis H_0; $\theta_1 = \sigma^2_{\alpha} = 0$, renders OLS the restricted MLE. Lejeune's second LM test considers a fixed effects one-way error components model where α_i is a fixed parameter to be estimated and the remainder error term is heteroskedastic with $\varepsilon_{it} \sim N\left(0, \sigma^2_{\varepsilon_{it}}\right)$ and $\sigma^2_{\varepsilon_{it}} = \sigma^2_{\varepsilon} h_{\varepsilon}\left(Z_{it}\theta_1\right)$. The joint hypothesis is H_0; $\alpha_i = \theta_1 = 0$ for all $i = 1, 2, .., N$. This again renders OLS to be the restricted MLE.

Holly and Gardiol (2000) derived a score test for homoskedasticity in a one-way error components model where the alternative model is that the α_i's are independent and distributed as $N(0, \sigma^2_{\alpha_i})$ where $\sigma^2_{\alpha_i} = \sigma^2_{\alpha} h_{\alpha}\left(F_i\theta_2\right)$. Here, F_i is a vector of p explanatory variables such that $F_i\theta_2$ does not contain a constant term and h_{α} is a strictly positive twice differentiable function satisfying $h_{\alpha}(0) = 1$ with $h'_{\alpha}(0) \neq 0$ and $h''_{\alpha}(0) \neq 0$. The score test statistic for H_0; $\theta_2 = 0$, turns out to be one half the explained sum of squares of the OLS regression of $(\hat{s}/\bar{s}) - \iota_N$ against the p regressors in F as in the Breusch and Pagan test for homoskedasticity. Here $\hat{s}_i = \hat{u}_i' \frac{J_T}{T} \hat{u}_i$ and $\bar{s} = \sum_{i=1}^{N} \hat{s}_i / N$ where $\widehat{u}$ denote the maximum likelihood residuals from the restricted model under H_0; $\theta_2 = 0$. This is a one-way homoskedastic error components model with $\alpha_i \sim N(0, \sigma^2_{\alpha})$.

In the spirit of the general heteroskedastic model of Randolph (1988) and Lejeune (1996), Baltagi, Bresson and Pirotte (2006) derived a *joint* Lagrange multiplier test for homoskedasticity, i.e., H_0; $\theta_1 = \theta_2 = 0$. Under the null hypothesis, the model is a homoskedastic one-way error components regression model. Note that this is different from Lejeune (1996), where under his null, $\sigma^2_{\alpha} = 0$. Allowing for $\sigma^2_{\alpha} > 0$ is more likely to be the case in panel data where heterogeneity across the individuals is likely to be present even if heteroskedasticity is not. The model under the null is exactly that of Holly and Gardiol (2000) but it is more general under the alternative since it does not assume a homoskedastic remainder error term. Next, Baltagi, et al. (2006) derived an LM test for the null hypothesis of homoskedasticity of the individual random effects assuming homoskedasticity of the remainder error term, i.e., $\theta_2 = 0 \mid \theta_1 = 0$. Not surprisingly, they get the Holly and Gardiol (2000) LM test. Last but not least, Baltagi et al. (2006) derived an LM test for the null hypothesis of homoskedasticity of the remainder error term assuming homoskedasticity of the individual effects, i.e., $\theta_1 = 0 \mid \theta_2 = 0$. Monte Carlo experiments showed that the joint LM test performed well when both error components were heteroskedastic, and performed second best when one of the components was homoskedastic while the other was not. In contrast, the marginal LM tests performed best when heteroskedasticity was present in the right error component. They yielded misleading results if heteroskedasticity was present in the wrong error component.

3.5 Estimation Using Unbalanced Panels

The presentation of the estimation and testing methods made above was assuming a balanced sample, i.e. that all individuals in the sample are observed over the same period of time. However, in practice, this is almost never the case. Some individuals disappear from the sample, others come in, some of them are absent at some dates, etc. Fortunately, all the methods above still apply with an unbalanced panel sets with only minor changes. As an illustration, let us consider the Feasible-GLS estimator. Following Baltagi (1985, 2005), one can write the corresponding regression as

$$y_{it} + (\sqrt{\widehat{\theta}_i} - 1)\overline{y}_i = [x'_{it} + (\sqrt{\widehat{\theta}_i} - 1)\overline{x}'_i]\,\beta + u_{it} + (\sqrt{\widehat{\theta}_i} - 1)\overline{u}_i \tag{3.20}$$

where

$$\widehat{\theta}_i = \frac{\widehat{\sigma}^2_\varepsilon}{\widehat{\sigma}^2_\varepsilon + T_i\widehat{\sigma}^2_\alpha}.$$

Then, the model transformation depends on the number T_i of observations of each individual i. It is then no more possible to estimate the variances σ^2_ε and σ^2_α directly from the Within and Between regressions.
However, the estimated residual variance from the Within regression, given by

$$\widehat{\sigma}^2_w = \frac{1}{\sum_{i=1}^{N} T_i - N - k_w} \sum_{i=1}^{N}\sum_{t=1}^{T} [(y_{it} - \overline{y}_i) - (x'_{it} - \overline{x}'_i)\widehat{\beta}]^2$$

still provides an unbiased and consistent estimate of σ^2_ε. On the contrary, it is no more the case for the Between regression residual variance. The reason is that the Between regression now relies on the individual means computed over T_i observations, which makes its disturbances become heteroscedastic. Indeed, we have

$$\overline{y}_i = \sum_{k=1}^{K} \beta_k \overline{x}_{ki} + \overline{u}_i \qquad \text{where } \overline{u}_i = \alpha_i + \overline{\varepsilon}_i, \quad i = 1, \ldots, N \tag{3.21}$$

with:

$$E(\overline{u}_i) = 0 \text{ but } V(\overline{u}_i) = \sigma^2_\alpha + \sigma^2_\varepsilon / T_i\,.$$

The consequence of this heteroscedasticity is the (obvious) inconsistency of the residual variance as an estimate of $\sigma^2_\alpha + \sigma^2_\varepsilon/T_i$ .However, the Between estimator of the coefficients (β) is still unbiased and consistent in this case and thus, one can consistently estimate σ^2_α by:

$$\widehat{\sigma}^2_\alpha = \frac{1}{N - (K+1)} \sum_{i=1}^{N} [(\overline{y}_i - \overline{x}'_i\widehat{\beta}_B)^2 - \frac{1}{T_i}\widehat{\sigma}^2_w].$$

Other ways to proceed by generalizing the procedures proposed by Wallace and Hussain (1969), Amemiya (1971), Swamy and Arora (1972) along the lines

suggested in Baltagi and Chang (1994) can be adopted as an alternative. Then, once estimates of $\widehat{\sigma}_{\alpha}^2$ and $\widehat{\sigma}_{\varepsilon}^2$ are obtained, it is easy to transform the model as described above to get the Feasible-GLS estimator.

References

Amemiya T., 1971, The estimation of the variances in a variance-components model, International Economic Review 12, 1–13.

Arellano M., 1987, Computing robust standard errors for within-groups estimators, Oxford Bulletin of Economics and Statistics 49, 431–434.

Baillie R.T. and Baltagi B.H., 1999, Prediction from the regression model with one-way error components, Chap. 10 in C. Hsiao, K. Lahiri, L.F. Lee and H. Pesaran eds., Analysis of Panels and Limited Dependent Variable Models (Cambridge University Press, Cambridge), 255–267.

Baltagi B.H., 1985, Pooling cross-sections with unequal time-series lengths, Economics Letters 18, 133–136.

Baltagi B.H., 1986, Pooling under misspecification: Some Monte Carlo evidence on the Kmenta and the error components techniques, Econometric Theory 2, 429–440.

Baltagi B.H., 1988, An alternative heteroscedastic error component model, problem 88.2.2, Econometric Theory 4, 349–350.

Baltagi B.H., 2005, Econometric Analysis of Panel Data (John Wiley, Chichester).

Baltagi B.H. and Chang Y.J., 1994, Incomplete panels: A comparative study of alternative estimators for the unbalanced one-way error component regression model, Journal of Econometrics 62, 67–89.

Baltagi B.H. and Griffin J.M., 1988, A generalized error component model with heteroscedastic disturbances, International Economic Review 29, 745–753.

Baltagi B.H. and Li Q., 1991a, A transformation that will circumvent the problem of autocorrelation in an error component model, Journal of Econometrics 48, 385–393.

Baltagi B.H. and Li Q., 1991b, Variance component estimation under misspecification, problem 91.3.3, Econometric Theory 7, 418–419.

Baltagi B.H. and Li Q., 1995, Testing AR (1) against MA (1) disturbances in an error component model, Journal of Econometrics 68, 133–151.

Baltagi B.H. and Li Q., 1997, Monte Carlo results on pure and pretest estimators of an error component model with autocorrelated disturbances, Annales D'Économie et de Statistique 48, 69–82.

Baltagi B.H., Bresson G. and Pirotte A., 2005, Adaptive estimation of heteroskedastic error component models, Econometric Reviews 24, 39–58.

Baltagi B.H., Bresson G. and Pirotte A., 2006, Joint LM test for Heteroskedasticity in a one-way error component model, Journal of Econometrics, 134, 401–417.

Beck N. and Katz, J., 1995, What to do (and not to do) with time-series-cross-section data in comparative politics, American Political Science Review 89, 634–647.

Breusch T.S., 1987, Maximum likelihood estimation of random effects models, Journal of Econometrics 36, 383–389.

Breusch T.S. and Godfrey L.G., 1981, A review of recent work on testing for autocorrelation in dynamic simultaneous models, in D.A. Currie, R. Nobay and D. Peel, eds., Macroeconomic Analysis, Essays in Macroeconomics and Economics (Croom Helm, London), 63–100.

Breusch T.S. and Pagan A.R., 1979, A simple test for heteroskedasticity and random coefficient variation, Econometrica 47, 1287–1294.

Breusch T.S. and Pagan A.R., 1980, The Lagrange multiplier test and its applications to model specification in econometrics, Review of Economic Studies 47, 239–253.

Burke S.P., Godfrey L.G. and Termayne A.R., 1990, Testing AR(1) against MA(1) disturbances in the linear regression model: An alternative procedure, Review of Economic Studies 57, 135–145.
Deschamps P., 1991, On the estimated variances of regression coefficients in misspecified error components models, Econometric Theory 7, 369–384.
Fuller W.A. and Battese G.E., 1974, Estimation of linear models with cross-error structure, Journal of Econometrics 2, 67–78.
Goldberger A.S., 1962, Best linear unbiased prediction in the generalized linear regression model, Journal of the American Statistical Association 57, 369–375.
Hausman J.A., 1978, Specification tests in econometrics, Econometrica 46, 1251–1271.
Hausman J.A. and Taylor W.E., 1981, Panel data and unobservable individual effects, Econometrica 49, 1377–1398.
Holly A. and Gardiol, L., 2000, A score test for individual heteroscedasticity in a one-way error components model, Chap. 10 in J. Krishnakumar and E. Ronchetti, eds., Panel Data Econometrics: Future Directions (North-Holland, Amsterdam), 199–211.
Honda Y., 1985, Testing the error components model with non-normal disturbances, Review of Economic Studies 52, 681–690.
Hsiao C., 1986, Analysis of Panel Data (Cambridge University Press, Cambridge).
Kmenta J., 1986, Elements of Econometrics (MacMillan, New York).
Lejeune B., 1996, A full heteroscedastic one-way error components model for incomplete panel: Maximum likelihood estimation and Lagrange multiplier testing, CORE discussion paper 9606, Universite Catholique de Louvain, 1–28.
Li Q. and Stengos T., 1994, Adaptive estimation in the panel data error component model with heteroskedasticity of unknown form, International Economic Review 35, 981–1000.
Lillard L.A. and Willis R.J., 1978, Dynamic aspects of earning mobility, Econometrica 46, 985–1012.
Maddala G.S., 1971, The use of variance components models in pooling cross section and time series data, Econometrica 39, 341–358.
Maddala G.S. and Mount T.D., 1973, A comparative study of alternative estimators for variance components models used in econometric applications, Journal of the American Statistical Association 68, 324–328.
Mazodier P. and Trognon A., 1978, Heteroskedasticity and stratification in error components models, Annales de l'INSEE 30-31, 451–482.
Mundlak Y., 1978, On the pooling of time series and cross-section data, Econometrica 46, 69–85.
Nerlove M., 1971b, A note on error components models, Econometrica 39, 383–396.
Phillips R.L., 2003, Estimation of a stratified error components model, International Economic Review 44, 501–521.
Prucha I.R., 1984, On the asymptotic efficiency of feasible Aitken estimators for seemingly unrelated regression models with error components, Econometrica 52, 203–207.
Randolph W.C., 1988, A transformation for heteroscedastic error components regression models, Economics Letters 27, 349–354.
Rao C.R., 1970, Estimation of heteroscedastic variances in linear models, Journal of the American Statistical Association 65, 161–172.
Rao C.R., 1972, Estimation variance and covariance components in linear models, Journal of the American Statistical Association 67, 112–115.
Rao S.R.S, Kaplan J. and Cochran W.C., 1981, Estimators for the one-way random effects model with unequal error variances, Journal of the American Statistical Association 76, 89–97.
Roy N., 2002, Is adaptive estimation useful for panel models with heteroscedasticity in the individual specific error component? Some Monte Carlo evidence, Econometric Reviews 21, 189–203.
Swamy P.A.V.B. and Arora, S.S., 1972, The exact finite sample proporties of the estimators of coefficients in the error components regression models, Econometrica 40, 253–260.
Taub A.J., 1979, Prediction in the context of the variance-components model, Journal of Econometrics 10, 103–108.

Taylor W.E., 1980, Small sample considerations in estimation from panel data, Journal of Econometrics 13, 203–223.
Trognon A., 1993, Econométrie des données individuelles-temporelles, Cours polycopié, ENSAE.
Verbon H.A.A., 1980, Testing for heteroscedasticity in a model of seemingly unrelated regression equations with variance components (SUREVC), Economics Letters 5, 149–153.
Wallace T.D. and Hussain A., 1969, The use of error components models in combining cross-section and time-series data, Econometrica 37, 55–72.
Wansbeek T.J., 1989, An alternative heteroscedastic error components model, solution 88.1.1, Econometric Theory 5, 326.

Chapter 4
Endogenous Regressors and Correlated Effects

Rachid Boumahdi and Alban Thomas

4.1 Introduction

There are several situations in econometric modeling where consistency of parameter estimates is questionable because some explanatory variables may be correlated with the model disturbances. Hence the fundamental exogeneity assumption for the regressors may not be supported by the data, with two implications. First, the source of this correlation might be investigated upon to propose possible corrections. Second, alternative but consistent estimators may be proposed.

One of the most well-known source of endogenous regressors is the case of simultaneous equations models, in which some of the regressors in a given equation are the dependent variables in others and consequently are correlated with the disturbances of the equation under consideration. Another cause of correlation between explanatory variables and model disturbances is when the former are subject to measurement errors. Chapter 9 provides a detailed treatment of simultaneity and measurement error issues in the case of panel data.

There is however an important reason why regressors may be endogenous in the context of panel data. As discussed in the preceding chapters, accounting for individual unobserved heterogeneity is usually done by incorporating random individual-specific effects to the usual idiosyncratic disturbances of the model. Consequently, regressors must be uncorrelated with these individual effects as well for consistent estimates to be obtained. This assumption of no-correlation has been widely criticized by many authors, among which Mundlak (1978).

Rachid Boumahdi
Toulouse School of Economics, GREMAQ and LIHRE; Université des Sciences Sociales de Toulouse, 21 Allée de Brienne, 31000 Toulouse, France, e-mail: boumahdi@univ-tlse1.fr

Alban Thomas
Toulouse School of Economics, INRA; Université des Sciences Sociales de Toulouse, 21 Allée de Brienne, 31000 Toulouse, France, e-mail: thomas@toulouse.inra.fr

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

Consider for example an agricultural production model (crop yield response function) where output depends on a set of inputs (labor, fertilizer, etc.). It is likely that variables outside the scope of farmer's decisions are also impacting the final crop output: soil characteristics (slope, water reserve, etc.) and climatic conditions. Land marginal productivity as represented by soil characteristics is often very difficult to observe with precision, and is often supposed to be part of the farm specific effect. But because farmer's input choice is likely to depend on land productivity, observed input levels are likely to be correlated with the farmer specific effect. This is especially true for fertilizer and water inputs, whose application levels are likely to be negatively correlated with systematic soil fertility and permanent water reserve, respectively.

Another popular example is the case of an individual earning function (wage equation), where the logarithm of the wage rate is explained by variables related to occupation, experience, and education. However, expected marginal productivity of a worker depends on individual ability, which is partly unobserved. In particular, individual ability may positively influence working wages, as well as education level of the individual. If the latter is an explanatory variable in the wage equation while being partly correlated with unobserved ability, individual effects (unobserved ability) may then be correlated with regressors.

This chapter addresses the issue of correlated effects, and endogenous regressors in the case of panel data. We present the main estimation and testing procedures employed in a single-equation, linear panel-data context. Starting with a brief overview of error structures and model transformations (fixed effects, first and quasi differences), we present Instrumental Variable (IV) and Generalized Method of Moments (GMM) procedures for consistent and efficient estimation of static models. We devote a particular section to augmented linear models with time-invariant regressors and show how to identify model parameters. Estimation of this kind of models with IV or GMM is discussed, and we compare in particular the efficiency of these estimators, depending on the validity of a no-conditional-heteroskedasticity assumption. A way to measure instrument relevance in the context of panel data models estimated by instrumental-variables procedures is presented, based on single-parameter information. Estimation by Instrumental Variable of models including time-varying regressors only is also the subject of a section, where endogenous regressors can be of any nature (time-varying only or not). As dynamic panel data models will be the subject of Chap. 8, we do not deal with the vast literature on the subject, that has emerged since the seminal work of Anderson and Hsiao (1982) and Arellano and Bond (1991). We conclude this chapter by a brief presentation of unbalanced panel data models with correlated effects and endogenous regressors, including nested error component models.

4.2 Estimation of Transformed Linear Panel Data Models

Consider the linear panel data model:

$$y_{it} = x_{it}\beta + u_{it}, \qquad i = 1,\ldots,N;\ t = 1,\ldots,T, \tag{4.1}$$

where x_{it}' is a $K \times 1$ vector regressors depending on individual i and time t except the first column of x_{it} which is a vector of ones. The error term u_{it} may contain unobserved individual heterogeneity components, as in the one-way error component specification, $u_{it} = \alpha_i + \varepsilon_{it}$. We assume for most of the chapter that the sample is balanced, i.e., each cross-sectional unit has the same number of non missing observations (T). The case of unbalanced panels will be briefly discussed in Sect. 4.7.

As discussed in Chaps. 2 and 3, a conditional (fixed effects) or a random effects approach will lead to similar results asymptotically under standard assumptions, among which exogeneity of the $x_{it}s$. On the other hand, when the correlation between u_{it} and some $x_{it}s$ in (4.1) is not accounted for, Ordinary or Generalized Least Squares estimators are not consistent. In this case, an easy way to cope with such endogeneity is simply to filter out this component. Such a strategy is applicable to a variety of error structures, as we now see.

4.2.1 Error Structures and Filtering Procedures

We present here basic transformations for eliminating the unobserved individual heterogeneity component in linear models. The motivation for such filtering in most cases comes from endogeneity issues, and in particular the fact that regressors are correlated with individual effects.

In most applications, the error component structure can be specified as a particular case of the following representation:

$$u_{it} = \alpha_i + \lambda_t v_i + \varepsilon_{it}, \tag{4.2}$$

where α_i and v_i are unobserved heterogeneity terms, λ_t is a time effect, and ε_{it} is i.i.d. across individuals and time periods. Let σ_α^2, σ_v^2 and σ_ε^2 respectively denote the variance of α_i, v_i and ε_{it}. The most important special cases are:

Case 1. (One-way error component model) $\lambda_t = \bar{\lambda}\ \forall t$.
Case 2. (Two-way error component model) $v_i = \bar{v}\ \forall i$.
Case 3. (Cross-sectional dependence Type I) $\alpha_i = \bar{\alpha}\ \forall i$.

Case 1 is by far the most widely used specification. When λ_t is constant across time periods, the error component structure reduces to $\alpha_i + \bar{\lambda} v_i + \varepsilon_{it} \equiv \alpha_i^* + \varepsilon_{it}$ (the one-way specification).

In case 2, λ_t can represent a trend function or simply consist of (non-monotonic) time effects that impact all units in a similar way for a given time period. It may however be of interest in applications to consider heterogeneous trends, where the marginal impact of the common time shock θ_t is individual-specific; this is obtained in case 3. In the general case of (4.2) where α_i and v_i are allowed to vary across units, we have both heterogeneous intercepts and slopes on the time effects.

Let us examine model transformations to eliminate heterogeneous individual heterogeneity terms in each of the cases presented above.

For case 1, the most common practice is to wipe out α_i with the Within-group (fixed effects) transformation, $\varepsilon_{it} - \bar{\varepsilon}_i = (y_{it} - \bar{y}_i) - (x_{it} - \bar{x}_i)\beta$, where $\bar{y}_i$ denotes the individual mean for unit i and variable y_{it}. This equation provides a simple way of obtaining consistent least squares estimation of β under the assumption of strong exogeneity: $E[(x_{it} - \bar{x}_i)|\varepsilon_{is}] = 0 \quad \forall s, \forall t$.

Alternatively, we may use the first-difference transformation $\Delta u_{it} = \Delta\varepsilon_{it} = \Delta y_{it} - \Delta x_{it}\beta = (y_{it} - y_{i,t-1}) - (x_{it} - x_{i,t-1})\beta$, and consistent estimation of β then obtains under the assumption that $E[\Delta x_{it}|\varepsilon_{it}, \varepsilon_{i,t-1}] = 0$, a somehow weaker assumption than above. In vector form, we can use the $T \times (T-1)$ submatrix L_T for performing first differences:

$$L_T = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & 0 & \cdots & 0 & -1 \end{bmatrix}.$$

Using first differences introduces a moving-average serial correlation on the transformed residual (Arellano and Bover, 1995). To remove such a correlation, it is possible to use the *Orthogonal deviation* procedure:

$$y_{it}^* = \frac{\sqrt{T-t}}{\sqrt{T-t+1}} \left[y_{it} - \frac{1}{T-t} \sum_{s=t+1}^{s=T} y_{is} \right], \tag{4.3}$$

$i = 1, \ldots, N \quad t = 1, \ldots, T-1$.

Whatever the transformation considered, be it within-group (fixed effects), first differences or orthogonal deviations, identification of parameter β is possible (except the constant term) because it is assumed that x_{it} is time-varying. First differences and deviations from individual means allow one to obtain the same information because operators Q_T (for fixed effects) and L_T (for first differences) span the same column space, with $Q_T = L_T(L_T'L_T)^{-1}L_T'$.

The choice between fixed effects and first differences, on the grounds of efficiency, depends in practice on assumptions made on homoskedasticity assumptions as follows. Maintaining the strict exogeneity assumption $E(\varepsilon_{it}|x_i, \alpha_i) = 0, t = 1, \ldots, T$, where $x_i = (x_{i1}, \ldots, x_{iT})$, if we further assume that $E(\varepsilon_i\varepsilon_i'|x_i, \alpha_i) = \sigma_\varepsilon^2 I_T$ (no heteroskedasticity nor serial correlation), then fixed effects is the most efficient estimator in the class of models satisfying these conditions. On the other hand, if we replace the latter assumption by

$$E(\Delta\varepsilon_i\Delta\varepsilon_i'|x_i, \alpha_i) = \sigma_{\Delta\varepsilon}^2 I_{T-1}, t = 2, \ldots, T,$$

then it can be shown that the first-difference estimator is more efficient. This is the case when ε_{it} follows a random walk.

In case 2, filtering of both individual and time effects can be achieved by means of a modified Within operator which simultaneously filters out time-invariant and

time-varying only components. We will discuss such transformation in detail in the section on time-varying only regressors below.

The model corresponding to case 3 was suggested by Holtz-Eakin, Newey and Rosen (1988), Ahn, Lee and Schmidt (2001), Lillard and Weiss (1979). Unless λ_t is constant across time periods, Within-group or first-difference transformations will fail to filter out the unobserved individual heterogeneity component α_i.

Define a new variable $r_t = \lambda_t/\lambda_{t-1}$; substracting from the equation at time t its expression lagged one period and premultiplied by r_t, we have

$$y_{it} - r_t y_{i,t-1} = (x_{it} - r_t x_{i,t-1})\beta + \varepsilon_{it} - r_t \varepsilon_{i,t-1}. \tag{4.4}$$

The transformed model using the *Quasi-differencing* technique is now a nonlinear equation with additional parameters to be estimated: $r_t, t = 2,3,\ldots,T$. Interestingly, parameters associated with time-invariant regressors become identified with a nonlinear regression of (4.4). This is the only case of such identification for those parameters in transformed models of the kind presented here.

Consider now the general case (4.1). To eliminate both effects α_i and v_i, it is necessary to use a double-transformation: first differences, and then quasi-differences:

$$\triangle y_{it} - \tilde{r}_t \triangle y_{i,t-1} = (\triangle x_{it} - \tilde{r}_t \triangle x_{i,t-1})\,\beta + \triangle \varepsilon_{it} - \tilde{r}_t \triangle \varepsilon_{i,t-1}, \tag{4.5}$$

$i = 1,2,\ldots,N, t = 3,4,\ldots,T$, where

$$\tilde{r}_t = \triangle\lambda_t/\triangle\lambda_{t-1} = (\lambda_t - \lambda_{t-1})/(\lambda_{t-1} - \lambda_{t-2}).$$

Such double transformation of the model has been suggested by Nauges and Thomas (2003) in the dynamic panel data context. Wansbeek and Knaap (1999) use a double first-difference transformation in the special case of a dynamic panel data model with a random trend with $\lambda_t = t$ (the random growth model, see Heckman and Holtz (1989)).

In what follows, we will mostly be working with the one-way error component model $u_{it} = \alpha_i + \varepsilon_{it}$.

4.2.2 An IV Representation of the Transformed Linear Model

Most estimators for linear panel data models can be shown to derive from the following orthogonality condition in matrix form:

$$E\left[A'(TU)\right] = 0 \quad \Leftrightarrow \frac{1}{N}A'TY = \frac{1}{N}A'TX\beta, \tag{4.6}$$

where A is a $NT \times L$ matrix of instruments and T is a $NT \times NT$ matrix transformation operator. Let $Q = I_{NT} - B$ and $B = I_N \otimes (1/T)e_T e_T'$ denote the Within and Between matrix operators respectively, where e_T is a T vector of ones. The fixed effects estimator obtains with $A = X$ and $T = Q$ so that $\hat{\beta}_W = (X'QX)^{-1}X'QY$ because Q is idempotent. In the one-way model, the GLS estimator obtains with $A = X$ and

$T = \Omega^{-1/2}$ so that $\hat{\beta}_{\text{GLS}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$, where the covariance matrix of $U = \alpha + \varepsilon$ is:

$$\Omega = \theta_1^2 Q + \theta_2^2 B,\ \Omega^{-1} = (1/\theta_1^2)Q + (1/\theta_2^2)B,\ \Omega^{-\frac{1}{2}} = (1/\theta_1)Q + (1/\theta_2)B, \quad (4.7)$$

where $\theta_1^2 = \sigma_\varepsilon^2,\ \theta_2^2 = \sigma_\varepsilon^2 + T\sigma_\alpha^2$.

Under the strict exogeneity assumption and assuming that the error structure is correctly represented, consistent estimates are obtained from moment conditions as in (4.6). Therefore, most popular estimators for linear panel data models can be represented in a IV form.

Depending on assumptions made on the error structure and the choice of the instrument matrix, estimators can be either inconsistent of inefficient, and it is therefore important to test for the validity of conditions underlying the construction of the estimator. To disentangle model misspecification due to an invalid set of instruments from an invalid transformation matrix, different specifications should be tested. Estimates constructed from either the same A but a different T, or the opposite, can be used to form a series of specification tests.

As presented in Chap. 3, the Generalized Least Squares (GLS) estimator may be selected on the grounds of efficiency in the case of the one-way linear panel data model, if assumptions underlying the random-effects specification are valid (in particular, strict exogeneity of the $x_{it}s$). If however, $E(\alpha_i x_{it}) \neq 0$, then GLS is not consistent, and fixed effects (or any transformation filtering out unobserved individual effects) should be used instead.

A very simple specification test is the Hausman exogeneity test, constructed as follows (Hausman, 1978). The null hypothesis to test is: $H_0 : E(x'_{it}\alpha_i) = 0 \forall i, \forall t$, and we have two estimators available. $\hat{\beta}_1$ (e.g., the GLS) is consistent are efficient under the null, and inconsistent otherwise, while the fixed effects estimator $\hat{\beta}_W$ is consistent under the null and under the alternative, but is not efficient (under the null).

The Hausman test for linear panel data is based on the fact that, under H_0, both estimators should be asymptotically equivalent, $\hat{\beta}_1$ being more efficient. The test statistic is

$$\text{HT} = \left(\hat{\beta}_W - \hat{\beta}_1\right)' \left[\text{Var}(\hat{\beta}_W) - \text{Var}(\hat{\beta}_1)\right]^{-1} \left(\hat{\beta}_W - \hat{\beta}_1\right) \backsim \chi^2(\tilde{K}),$$

where $\tilde{K}$ is the column dimension of $\hat{\beta}_W$. Note that $\hat{\beta}_1$ and $\hat{\beta}_W$ must have the same dimension, i.e., parameters identified with the fixed effects procedure. Also, the weighting matrix $\left[\text{Var}(\hat{\beta}_W) - \text{Var}(\hat{\beta}_1)\right]$ is always semidefinite positive because $\hat{\beta}_1$ is more efficient than Within under the null.

Finally, concerning the interpretation of the number of degrees of freedom of the test, the Within estimator is based on the condition $E(X'QU) = 0$, whereas $\hat{\beta}_1$ is based on a larger set of moment conditions. This is in fact the origin of the difference in efficiency between both estimators. In the case of GLS, the set of conditions is $E(X'^{-1}U) = 0 \Rightarrow E(X'QU) = 0$ and $E(X'BU) = 0$, and we therefore add K additional conditions (in terms of B), which is the rank of X.

It is important to note at this stage that both cases considered up to now are rather polar (extreme) cases: either all of the explanatory variables are endogenous, or neither of them is.

If we do not wish to maintain the assumption that all regressors are correlated with individual effects, an alternative estimation method may be considered: Two-Stage Least Squares (2SLS) or Instrumental Variable (IV) estimation. Recall that in a cross-section context with N observations, the model would be:

$$Y = X\beta + \varepsilon, \quad E(X'\varepsilon) \neq 0, \quad E(A'\varepsilon) = 0, \tag{4.8}$$

where A is a $N \times L$ matrix of instruments. If $K = L$, the orthogonality condition is

$$\left[A'(Y - X\beta)\right] = 0 \quad \Leftrightarrow \quad (A'Y) = (A'X)\beta, \tag{4.9}$$

and the IV estimator is $\hat{\beta} = (A'X)^{-1}A'Y$. If $L > K$, the model is over-identified (L conditions on K parameters). For any matrix A, let $P_{[A]} = A(A'A)^{-1}A'$ be the projection onto the column space of A. We can construct the quadratic form $(Y - X\beta)'P_{[A]}(Y - X\beta)$ and the IV estimator is $\hat{\beta} = (X'P_{[A]}X)^{-1}(X'P_{[A]}Y)$.

In the cross section context, instruments A originate outside the structural equation. In panel data models however, as we will see below, the advantage is that instruments (not correlated with the individual effect) can be obtained directly. Another important difference in practice is that, when dealing with panel data, spherical disturbances can no longer be assumed.

4.3 Estimation with Time-Invariant Regressors

4.3.1 Introduction

When considering estimation of a model with correlated effects, two arguments are in favor of yet another estimation procedure than Fixed Effects. First, one can sometimes obtain more efficient parameter estimates than the Within. Second, using the Within estimator does not enable us to estimate parameters associated to time-invariant explanatory variables. Indeed, as the estimator is built upon differentiating all variables with respect to individual means, then all variables which are individual-specific are dropped from the equation to be estimated.

For these reasons, an estimation method based on instrumental variables is called for. As we will show, Instrumental-Variables (IV) estimators yield more efficient estimators than the Within procedure, while allowing identification of all parameters in the model. To motivate its use, we are going to present in this section an augmented model, in which some of the explanatory variables may be endogenous, and some regressors are not time-varying but only individual-specific. Including individual-specific variables z_i is indeed important from an empirical perspective, as many samples contain important information on individuals, which does not vary

over time (e.g., sex, education completed, place of residence if individuals have not moved during the whole sample period).

Hausman and Taylor (1981) – hereafter HT – consider the following model:

$$y_{it} = x_{it}\beta + z_i\gamma + \alpha_i + \varepsilon_{it}, \qquad i = 1,\cdots,N;\ t = 1,\cdots,T, \tag{4.10}$$

where ε_{it} is assumed to be uncorrelated with x_{it}, z_i and α_i while the effects α_i may be correlated with some explanatory variables in x_{it} and/or z_i.

Stacking all NT observations we can write (4.10) as: $Y = X\beta + Z\gamma + \alpha + \varepsilon$, where Y is $NT \times 1$, X is $NT \times K$, Z is $NT \times G$, ε and α are $NT \times 1$ respectively. If X and Z are uncorrelated with α, the Generalized Least Squares (GLS) estimator yields consistent and efficient parameter estimates:

$$\hat{\mu}_{\text{GLS}} = \left[\frac{1}{\theta_1^2}\Phi' Q\Phi + \frac{1}{\theta_2^2}\Phi' B\Phi\right]^{-1}\left[\frac{1}{\theta_1^2}\Phi' QY + \frac{1}{\theta_2^2}\Phi' BY\right], \tag{4.11}$$

where $\Phi = [X, Z]$ and $\mu' = [\beta', \gamma']$. This estimator may generally be found more simply computationally by first transforming X, Z and Y to $Y^* = \Omega^{-\frac{1}{2}}Y$, $X^* = \Omega^{-\frac{1}{2}}X$ and $Z^* = \Omega^{-\frac{1}{2}}Z$ and then estimating β and γ from the Ordinary Least Squares (OLS) regression of Y^* on X^* and Z^*. The estimated variance–covariance matrix of the GLS estimator $\hat{\delta}_{\text{GLS}}$ is:

$$V(\hat{\mu}_{\text{GLS}}) = \hat{\sigma}_\varepsilon^2\left[\frac{1}{\hat{\theta}_1^2}\Phi' Q\Phi + \frac{1}{\hat{\theta}_2^2}\Phi' B\Phi\right]^{-1}, \tag{4.12}$$

where $\hat{\sigma}_\varepsilon^2 = \hat{\theta}_1^2 = \hat{u}_W'\hat{u}_W/(NT - K - G)$, $\hat{\theta}_2^2 = \hat{u}_B'\hat{u}_B/(N - K)$, $\hat{u}_W$ and $\hat{u}_B$ are the within and the between residual respectively.

4.3.2 Instrumental Variable Estimation

Following HT, we partition X and Z as follows:

$$X = [X_1, X_2] \text{ and } Z = [Z_1, Z_2],$$

where X_1 is $NT \times k_1$, X_2 is $NT \times k_2$, Z_1 is $NT \times g_1$ and Z_2 is $NT \times g_2$, so that the model in matrix form is

$$Y = X_1\beta_1 + X_2\beta_2 + Z_1\gamma_1 + Z_2\gamma_2 + \alpha + \varepsilon. \tag{4.13}$$

HT distinguish columns of X and Z which are asymptotically uncorrelated with α from those which are not. They assume, for fixed T and $N \to \infty$, that

$$\text{plim}\frac{1}{N}(BX_1)'\alpha = 0,\ \text{plim}\frac{1}{N}(BX_2)'\alpha \neq 0,\ \text{plim}\frac{1}{N}Z_1'\alpha = 0,\ \text{plim}\frac{1}{N}Z_2'\alpha \neq 0.$$

The way to estimate model (4.13) using an IV procedure is to rely on the exogeneity conditions above to construct a matrix of instruments. However, the method used differs from the standard one in simultaneous-equations literature. In the latter, a single equation is often estimated, which incorporates some endogenous variables among the regressors. All exogenous variables in the system are used as instruments, that is, exogenous variables not entering the equation of interest are also accounted for. In our case however, all the information is already contained in the single equation, meaning that we are able to construct instruments from variables in (4.13) alone. To see this, note that we are looking for instrument variables not correlated with the individual effect α. There are three ways such instruments may be found. First, exogenous variables X_1 and Z_1 are readily available because of the exogeneity conditions given above. Second, we may also obtain additional instruments through transformations of the original exogenous variables, because such transformations will also be exogenous. Third, we may consider as well transformations of endogenous variables, provided these transformations are not correlated with α.

An important aspect of panel data methods is that required transformations are very easily obtained through the use of matrices Q and B defined before. Matrix B calculates individual means of variables across all time periods, leaving the individual component unchanged. Therefore BX_1 is clearly applicable as an instrument, whereas BX_2 would not be, because endogeneity in X_2 comes through the individual component which is correlated with α. The Q matrix operates differentiation from individual means, filtering out the individual component. Therefore, QX_1 and QX_2 are also valid instruments, although the original X_2 variable is endogenous.

These considerations led HT to propose an IV estimator for a model corresponding to our (4.14). Their instrument matrix A^{HT} is the following:

$$A^{\mathrm{HT}} = (A_1^{\mathrm{HT}}, A_2^{\mathrm{HT}}),$$

where $A_1^{\mathrm{HT}} = (QX_1,\ QX_2)$ and $A_2^{\mathrm{HT}} = (BX_1,\ Z_1)$. We can show that:

$$P_{[A^{\mathrm{HT}}]} = A^{\mathrm{HT}}(A^{\mathrm{HT}\prime}A^{\mathrm{HT}})^{-1}A^{\mathrm{HT}\prime} = P_{[A_1^{\mathrm{HT}}]} + P_{[A_2^{\mathrm{HT}}]}. \tag{4.14}$$

To compute the efficient HT estimator we transform (4.13) by premultiplying it by $\Omega^{-\frac{1}{2}}$, so that the error term will have a diagonal covariance matrix. Using HT instruments $A_1^{\mathrm{HT}} = (QX_1,\ QX_2)$ and $A_2^{\mathrm{HT}} = (BX_1,\ Z_1)$, the IV estimator can be written as:

$$\hat{\mu}_{\mathrm{IV}} = \left[\frac{1}{\theta_1^2}\Phi' P_{[A_1^{\mathrm{HT}}]}\Phi + \frac{1}{\theta_2^2}\Phi' P_{[A_2^{\mathrm{HT}}]}\Phi\right]^{-1}\left[\frac{1}{\theta_1^2}\Phi' P_{[A_1^{\mathrm{HT}}]}Y + \frac{1}{\theta_2^2}\Phi' P_{[A_2^{\mathrm{HT}}]}Y\right], \tag{4.15}$$

and its variance–covariance matrix is

$$\mathrm{Var}\,(\hat{\mu}_{\mathrm{IV}}) = \hat{\sigma}_\varepsilon^2\left[\frac{1}{\theta_1^2}\Phi' P_{[A_1^{\mathrm{HT}}]}\Phi + \frac{1}{\theta_2^2}\Phi' P_{[A_2^{\mathrm{HT}}]}\Phi\right]^{-1}. \tag{4.16}$$

Breusch, Mizon and Schmidt (1989) -hereafter BMS- show that this is equivalent to using the alternative instrument matrices A^{HT}, C^{HT} and D^{HT} defined as follows

$$\begin{array}{rlll} A^{\mathrm{HT}} = (A_1^{\mathrm{HT}}, & A_2^{\mathrm{HT}}), & A_1^{\mathrm{HT}} = (QX_1, QX_2), & A_2^{\mathrm{HT}} = (BX_1, Z_1) \\ \text{or } C^{\mathrm{HT}} = (C_1^{\mathrm{HT}}, & C_2^{\mathrm{HT}}), & C_1^{\mathrm{HT}} = (Q), & C_2^{\mathrm{HT}} = (X_1, Z_1) \\ \text{or } D^{\mathrm{HT}} = (D_1^{\mathrm{HT}}, & D_2^{\mathrm{HT}}), & D_1^{\mathrm{HT}} = (QX_1, QX_2), & D_2^{\mathrm{HT}} = (X_1, Z_1). \end{array}$$

We will not enter into too much detail about these equivalences (see BMS, 1989 for more). Note however that the superiority of IV over Within estimators is easily seen, as far as the estimation of parameters β is concerned. The fixed effects procedure amounts to using the Q matrix as a single instrument. As it is well known that an IV estimator is more efficient when we add instruments, it is clear that the Hausman–Taylor estimator is more efficient than the Within estimator, since it entails (BX_1, Z_1) as additional instruments.

A final difficulty with IV estimators concerns estimation of variance components, because endogeneity of some regressors will yield inconsistent estimates of σ_α^2 and σ_ε^2 if the standard Feasible GLS procedure is used. Hausman and Taylor (1981) describe a method for obtaining consistent estimates. Let $\hat{\eta}$ denote the Within residual averaged over time periods:

$$\begin{aligned} \hat{\eta} &= BY - BX\hat{\beta}_W = (B - BX(X'QX)^{-1}X'Q)Y \\ &= Z\gamma + \alpha + B\varepsilon - BX(X'QX)^{-1}X'Q\varepsilon. \end{aligned} \tag{4.17}$$

If the last three terms in the equation above are treated as zero-mean residuals, then OLS and GLS estimates of γ will be inconsistent. However, consistent estimation is possible if the columns of X_1 provide sufficient instruments for the columns of Z_2. A necessary condition is that $k_1 \geqslant g_2$. The IV estimator of γ is

$$\hat{\gamma}_B = \left[Z'P_{[R]}Z\right]^{-1}\left[Z'P_{[R]}\hat{\eta}\right], \tag{4.18}$$

where $R = (X_1, Z_1)$. Now, using parameters estimates $\hat{\beta}_W$ and $\hat{\gamma}_B$, one forms the residuals

$$\hat{u}_W = QY - QX\hat{\beta}_W \quad \text{and} \quad \hat{u}_B = BY - BX\hat{\beta}_W - Z\hat{\gamma}_B. \tag{4.19}$$

These two vectors of residuals are finally used in the computation of the variance components as follows.[1]

$$\hat{\sigma}_\varepsilon^2 = \frac{\hat{u}_W'\hat{u}_W}{NT - N} \quad \text{and} \quad \hat{\sigma}_\alpha^2 = \frac{\hat{u}_B'\hat{u}_B}{N} - \frac{1}{T}\hat{\sigma}_\varepsilon^2$$

4.3.3 More Efficient IV Procedures

The Hausman–Taylor IV procedure has proved very popular, because of its relative computational simplicity and intuitive appeal. Since then however, there has been

[1] For details, see Hausman and Taylor (1981), p. 1384.

several improvements along its lines which led to more efficient estimation procedures.

The instruments used by Hausman and Taylor require only minimal exogeneity assumptions on variables, i.e., BX_1 and Z_1 are not correlated with the individual effect. As a consequence, this estimator may not be the most efficient if exogeneity conditions can be made more restrictive. Amemiya and MaCurdy (1986) – hereafter AM – suggested a potentially more efficient estimator by assuming that realizations of X_1 are not correlated with α in each time period, i.e., for all $t = 1,\ldots,T$ and $N \to \infty$ they assume that $\text{plim}(1/N)x'_{1it}\alpha_i = 0$. Consequently, we may not only use BX_1 as an instrument for individual i at time t, but also the whole series $(x_{1,i1}, x_{1,i2}, \ldots, x_{1,iT})$. AM define the following $NT \times Tk_1$ matrix:

$$X_1^* = \text{vec}\left\{e_T \otimes x'_{1,i}\right\} = \left\{e_T \otimes x'_{1,1}, \ldots, e_T \otimes x'_{1,N}\right\}, \quad \text{where } x_{1,i} = (x_{1,i1}, \ldots, x_{1,iT})',$$

which is such that $QX_1^* = 0$ and $BX_1^* = X_1^*$. Their instrument matrix is $A^{\text{AM}} = (A_1^{\text{AM}}, A_2^{\text{AM}})$, where $A_1^{\text{AM}} = (QX_1, QX_2)$ and $A_2^{\text{AM}} = (X_1^*, Z_1)$. An equivalent estimator obtains by using the matrix $C^{\text{AM}} = (C_1^{\text{AM}}, C_2^{\text{AM}})$, where $C_1^{\text{AM}} = (QX_1, QX_2)$ and $C_2^{\text{AM}} = [(QX_1)^*, BX_1, Z_1]$, $(QX_1)^*$ is constructed the same way as X_1^* above.

These authors suggest that their estimator is at least as efficient as Hausman–Taylor if individual effects are not correlated with regressors X_1 for each time period.

Note that the AM estimator differs from HT estimator only in its treatment of X_1. In fact, $A_1^{\text{HT}} = A_1^{\text{AM}}$ and $C_2^{\text{AM}} = ((QX_1)^*, BX_1, Z_1)$ differs from $A_2^{\text{HT}} = (BX_1, Z_1)$ only by using $(QX_1)^*$. In other words, HT use X_1 as two instruments namely QX_1 and BX_1 whereas AM use each such variable as $T+1$ instruments: $(QX_1)^*$ and BX_1.

Finally, a third IV method was described in BMS. Following these authors, if the variables in X_2 are correlated with effects only through a time-invariant component, then (QX_2) would not contain this component and $(QX_2)^*$ is a valid instrument. Their estimator is thus based on the following instrument matrix : $A^{\text{BMS}} = (A_1^{\text{BMS}}, A_2^{\text{BMS}})$, where $A_1^{\text{BMS}} = (QX_1, QX_2)$ and $A_2^{\text{BMS}} = [(QX_1)^*, (QX_2)^*, BX_1, Z_1]$. The estimated variance–covariance matrix of the IV estimator $\hat{\hat{\delta}}_{\text{IV}}$ has the same form as in (4.16), where $\hat{\sigma}_u^2 = \hat{u}'_{\text{IV}}\hat{u}_{\text{IV}}/(NT - K - G)$ and $\hat{u}_{\text{IV}}$ is the IV residual.

The Hausman test statistic can be used to check for the vality of the alternative IV estimators described above. The HT-IV estimator can first be compared with the fixed effects, to check that exogeneity assumption on X_1 and Z_1 are valid? If this is the case, then the more efficient procedures of AM-IV and BMS-IV can be compared with HT-IV to check that additional assumptions described above are supported by the data. See Cornwell and Rupert (1988) for an illustration of these test procedures.

4.4 A Measure of Instrument Relevance

It may be interesting in practice to investigate the performance of instruments in terms of efficiency of IV estimators on an individual-regressor basis. Cornwell and Rupert (1988) and Baltagi and Khanti-Akom (1990) have investigated efficiency

gains of instrumental variable estimators by fitting a wage equation on panel data and applying the methods proposed by HT, AM and BMS. Cornwell and Rupert (1988) found that efficiency gains are limited to the coefficient of time-invariant endogenous variables Z_2.

However, Baltagi and Khanti-Akom (1990) using the canonical correlation coefficient for comparing different sets of instrumental variables found that efficiency gains are not limited to the time-invariant variable. They also show that the geometric average of canonical correlations increases as one moves from HT to AM, and then from AM to BMS. In fact, the canonical correlations only measure instrument relevance for the group of endogenous regressors taken as a whole, but cannot be used to measure how a particular group of instruments affects relevance for one endogenous regressor as opposed to another.

More recently, Boumahdi and Thomas (2006) have extended the method proposed by Shea (1997) and Godfrey (1999) to the case of panel data. This method allows for measuring instrument relevance for separate endogenous regressors. Following Shea (1997) and Godfrey (1999), we consider estimation of a single parameter by rewriting the augmented model $Y = X\beta + Z\gamma + \alpha + \varepsilon$ as

$$Y = M\delta + \alpha + \varepsilon = M_1\delta_1 + M_2\delta_2 + \varepsilon, \tag{4.20}$$

where $M = [X, Z]$ and $\delta' = [\beta', \gamma']$, M_1 is $NT \times 1$ and M_2 is $NT \times (K + G - 1)$.

Define $\underline{M}_1 = (I_{NT} - P_{M_2})M_1$, $\overline{M}_1 = (I_{NT} - P_{\hat{M}_2})\hat{M}_1$ and $\hat{M}_j = P_A M_j$, $j = 1, 2$ where A is the matrix of instruments. In our panel data model, δ_1 would for example correspond to the first variable in $\Omega^{-\frac{1}{2}}X_2$. These definitions imply that $\overline{M}_1'\underline{M}_1 = \overline{M}_1'\overline{M}_1$. Using the same idea as in Shea (1997) and Godfrey (1999) in the case of a linear multiple regression model, we can use as a measure of instrumental variable relevance, the population squared correlation between $\overline{M}_1$ and $\underline{M}_1$ for the model:

$$\rho_p^2 = \operatorname{plim} \frac{\left(\overline{M}_1'\underline{M}_1\right)^2}{\left(\overline{M}_1'\overline{M}_1\right)\left(\underline{M}_1'\underline{M}_1\right)} = \operatorname{plim} \frac{\overline{M}_1'\overline{M}_1}{\underline{M}_1'\underline{M}_1}. \tag{4.21}$$

In applied work, provided N tends to infinity, we can approximate plim $\overline{M}_1'\overline{M}_1 / \underline{M}_1'\underline{M}_1$ by the following coefficient

$$R_p^2 = \frac{\overline{M}_1'\overline{M}_1}{\underline{M}_1'\underline{M}_1}. \tag{4.22}$$

It is not necessary in practice to compute the above expression, because the coefficient R_p^2 is directly related to the estimated parameter standard errors. To see this, consider the estimated variance of the first component of $\widehat{\delta}_{\mathrm{GLS}}$ and the corresponding component in $\widehat{\delta}_{\mathrm{IV}}$:

$$V\left(\widehat{\delta}_1^{\mathrm{GLS}}\right) = \widehat{\sigma}_{\eta_{\mathrm{GLS}}}^2 \left(\underline{M}_1'\underline{M}_1\right)^{-1}, \quad V\left(\widehat{\delta}_1^{\mathrm{IV}}\right) = \widehat{\sigma}_{\eta_{\mathrm{IV}}}^2 \left(\overline{M}_1'\overline{M}_1\right)^{-1}.$$

Then, R_p^2 can be written as

$$R_p^2 = \frac{\widehat{\sigma}^2_{\eta_{\text{IV}}} V\left(\widehat{\delta}_1^{\text{GLS}}\right)}{\widehat{\sigma}^2_{\eta_{\text{GLS}}} V\left(\widehat{\delta}_1^{\text{IV}}\right)} = \frac{\overline{M}_1' \overline{M}_1}{\underline{M}_1' \underline{M}_1}. \tag{4.23}$$

Consequently, the measure of instrumental variable relevance can be directly obtained by inspecting individual parameter (squared) standard errors.

4.5 Incorporating Time-Varying Regressors

Wyhowski (1994), Boumahdi and Thomas (1997) have extended the augmented model by incorporating time-varying regressors, i.e., variables which are not individual-specific, only time-period-specific. Think for example of a wage equation depending on individual-specific variables such as sex and education, and on time-varying regressors such as unemployment rate, economy-wide growth rate, etc. The intuition behind such a model would be that all individuals are affected by macro-economic variables the same way on average. Consider the two-way error component model as case 2 defined above:

$$u_{it} = \alpha_i + \lambda_t + \varepsilon_{it}. \tag{4.24}$$

The extended model we are considering is now the following:

$$y_{it} = x_{it}\beta + z_i\gamma + w_t\delta + \alpha_i + \lambda_t + \varepsilon_{it}, \qquad i = 1, \cdots, N;\ t = 1, \cdots, T, \tag{4.25}$$

where x_{it}' is a $K \times 1$ vector of time-varying explanatory variables, z_i' is a $G \times 1$ vector of time-invariant explanatory variables, and w_t' is a $H \times 1$ vector of individual-invariant explanatory variables. Unobserved effects α_i and λ_t are assumed to have zero mean and variances σ_α^2 and σ_λ^2 respectively. We assume further that $E(\varepsilon_{it}) = 0$, $E(\varepsilon_{it}\varepsilon_{is}) = \sigma_\varepsilon^2$ for $t = s$, $E(\varepsilon_{it}\varepsilon_{is}) = 0$ otherwise and $E(\alpha_i\varepsilon_{it}) = E(\lambda_t\varepsilon_{it}) = 0 \forall i, \forall t$. Stacking all NT observations we can write the model in a compact form as:

$$Y = X\beta + Z\gamma + W\delta + \alpha + \lambda + \varepsilon. \tag{4.26}$$

Let us introduce some notation for this model. As before, B is the Between matrix transforming variables into their means across periods (individual means); we now define $\bar{B}$ as a matrix transforming a variable into its mean across individuals (time mean). Hence, BY is time-invariant and individual-specific, whereas $\bar{B}$ is time-varying and independent from individuals. Let

$$B = I_N \otimes \frac{1}{T} e_T e_T', \qquad \bar{B} = \frac{1}{N} e_N e_N' \otimes I_T,$$

$$Q = I_{NT} - B - \bar{B} + J, \qquad J = \frac{1}{NT} e_{NT} e_{NT}' = B\bar{B}.$$

The new matrix Q allows to differentiate a given variable according to both its time and individual means. The J operator performs the total mean of a variable, i.e., JX is a $NT \times 1$ matrix with the same argument $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}$. With this notation, the variance–covariance matrix of the error term U reads:

$$\Omega = \theta_1^2 S_1 + \theta_2^2 S_2 + \theta_3^2 S_3 + \theta_4^2 J, \tag{4.27}$$

where $\theta_1^2 = \sigma_\varepsilon^2$, $\theta_2^2 = \sigma_\varepsilon^2 + T\sigma_\alpha^2$, $\theta_3^2 = \sigma_\varepsilon^2 + N\sigma_\lambda^2$, $\theta_4^2 = \sigma_\varepsilon^2 + N\sigma_\lambda^2 + T\sigma_\alpha^2$, $S_1 = I_{TN} - S_2 - S_3 - J$, $S_2 = B - J$, $S_3 = \bar{B} - J$, $S_k S_l = 0$ and $JJ' = J$ for $l \neq k$ and $k,l = 1,2,3$.

It is easy to show that[2]

$$\Omega^{-1} = (1/\theta_1^2)S_1 + (1/\theta_2^2)S_2 + (1/\theta_3^2)S_3 + (1/\theta_4^2)J, \tag{4.28}$$

and

$$\Omega^{-\frac{1}{2}} = (1/\theta_1)S_1 + (1/\theta_2)S_2 + (1/\theta_3)S_3 + (1/\theta_4)J. \tag{4.29}$$

If we assume that X, Z and W are uncorrelated with α and λ, then model parameters can be estimated by GLS as follows:

$$\hat{\nu}_{\text{GLS}} = \left[\sum_{k=1}^{3} \frac{1}{\theta_k^2}\Psi' S_k \Psi\right]^{-1} \left[\sum_{k=1}^{3} \frac{1}{\theta_k^2}\Psi' S_k Y\right], \tag{4.30}$$

where $\Psi = [X, Z, W]$ and $\nu' = [\beta', \gamma', \delta']$.

4.5.1 Instrumental Variables Estimation

Following HT and Wyhowski (1994) we allow for correlation between a subset of (X, Z, W) and (α, λ), and we partition X, Z and W as follows:

$$X = (X_1,\ X_2,\ X_3,\ X_4),\ Z = (Z_1,\ Z_2) \text{ and } W = (W_1,\ W_2).$$

Their dimensions are denoted as follows: $k_1, k_2, k_3, k_4, g_1, g_2, h_1$ and h_2 for X_1, X_2, X_3, Z_1, Z_2, W_1 and W_2 respectively. Furthermore, we assume that X_1 is not correlated with α and λ, X_2 is correlated with α but not λ, X_3 is correlated with λ but not α, X_4 is correlated with both λ and α.[3]

However, Z_1 and W_1 are assumed uncorrelated with α and λ respectively. In other words and following Wyhowski (1994), we assume that, for T fixed and $N \longrightarrow \infty$:

$$\text{plim}\,(S_2 X_1)'\alpha = 0,\ \text{plim}\,(S_2 X_3)'\alpha = 0,\ \text{plim}\,(S_2 Z_1)'\alpha = 0,$$

and, for N fixed and $T \longrightarrow \infty$:

$$\text{plim}\,(S_3 X_1)'\lambda = 0,\ \text{plim}\,(S_3 X_2)'\lambda = 0,\ \text{plim}\,(S_3 W_1)'\lambda = 0.$$

[2] We can also show that $\Omega = \sigma_\varepsilon^2 I_{NT} + T\sigma_\alpha^2 + N\sigma_\lambda^2$.

[3] Boumahdi and Thomas (1997) have considered another partition of X, Z and W.

Under this assumption, we can use as an appropriate instrument set:

$$A^{\mathrm{HT}} = (A_1^{\mathrm{HT}}, A_2^{\mathrm{HT}}, A_3^{\mathrm{HT}}),$$

where $A_1^{\mathrm{HT}} = (S_1 X)$, $A_2^{\mathrm{HT}} = (S_2 X_1, S_2 X_3, S_2 Z_1)$ and $A_2^{\mathrm{HT}} = (S_3 X_1, S_3 X_2, S_3 W_1)$.

Then the HT estimator can be written as:

$$\hat{\nu}_{\mathrm{IV}} = \left[\sum_{k=1}^{3} \frac{1}{\theta_k^2} \Psi' P_{[A_k^{\mathrm{HT}}]} \Psi \right]^{-1} \left[\sum_{k=1}^{3} \frac{1}{\theta_k^2} \Psi' P_{[A_k^{\mathrm{HT}}]} Y \right], \tag{4.31}$$

where A_k^{HT} is the matrix of instruments. The order condition for existence of the estimator can be obtained by counting instruments and parameters to be estimated. For parameters γ we must have:

$$K + k_1 + k_3 + g_1 \geq K + G \text{ or } k_1 + k_3 \geq g_2,$$

and for parameters δ, we must have:

$$K + k_2 + h_1 \geq K + H \text{ or } k_1 + k_2 \geq h_2,$$

where $K = k_1 + k_2 + k_3 + k_4$, $G = g_1 + g_2$ and $H = h_1 + h_2$.

Now, if we assume that $\operatorname{plim}(S_2 X_1)'\alpha = 0$ and $\operatorname{plim}(S_2 X_3)'\alpha = 0$, $\forall t = 1, \ldots, T$, and following AM, X_1 and X_3 can be used as two instruments: $(S_1 X_1, S_1 X_3)$ and X_1^*, X_3^*. X_1^* is the $NT \times Tk_1$ matrix defined as in the one-way AM case presented above, and

$$X_3^* = \operatorname{vec}\left\{ e_T \otimes x'_{3,i} \right\} = \left\{ e_T \otimes x'_{3,1}, \ldots, e_T \otimes x'_{3,N} \right\}, \quad \text{where } x_{3,i} = (x_{3,i1}, \ldots, x_{3,iT})'.$$

Furthermore, if we assume that $\operatorname{plim}(S_3 X_1)'\lambda = 0$ and $\operatorname{plim}(S_3 X_2)'\lambda = 0$ for each $i, i = 1, \ldots, N$, then X_1 and X_2 can be used as two instruments $(S_1 X_1, S_1 X_2)$ and X_1^0, X_3^0, where

$$X_1^0 = \begin{bmatrix} X_{1,11} & X_{1,21} & \ldots & X_{1,N1} \\ \vdots & \vdots & \vdots & \vdots \\ X_{1,1T} & X_{1,2T} & \ldots & X_{1,NT} \\ X_{1,11} & X_{1,21} & \ldots & X_{1,N1} \\ \vdots & \vdots & \vdots & \vdots \\ X_{1,1T} & X_{1,21} & \ldots & X_{1,N1} \\ \ldots & \ldots & \ldots & \ldots \\ \vdots & \vdots & \vdots & \vdots \\ \ldots & \ldots & \ldots & \ldots \\ X_{1,11} & X_{1,21} & \ldots & X_{1,N1} \\ \vdots & \vdots & \vdots & \vdots \\ X_{1,1T} & X_{1,2T} & \ldots & X_{1,NT} \end{bmatrix} \text{ and } X_2^0 = \begin{bmatrix} x_{2,11} & x_{2,21} & \ldots & x_{2,N1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{2,1T} & x_{2,2T} & \ldots & x_{2,NT} \\ x_{2,11} & x_{2,21} & \ldots & x_{2,N1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{2,1T} & x_{2,21} & \ldots & x_{2,N1} \\ \ldots & \ldots & \ldots & \ldots \\ \vdots & \vdots & \vdots & \vdots \\ \ldots & \ldots & \ldots & \ldots \\ x_{2,11} & x_{2,21} & \ldots & x_{2,N1} \\ \vdots & \vdots & \vdots & \vdots \\ x_{2,1T} & x_{2,2T} & \ldots & x_{2,NT} \end{bmatrix}.$$

In this case, the AM instruments can be defined as follows:

$$A^{\mathrm{AM}} = (A_1^{\mathrm{AM}}, A_2^{\mathrm{AM}}, A_3^{\mathrm{AM}}),$$

where $A_1^{\mathrm{AM}} = (S_1 X) = A_1^{\mathrm{HT}}, A_2^{\mathrm{AM}} = [A_2^{\mathrm{HT}}, (\bar{S}X_1, \bar{S}X_3)^*]$ and $A_3^{\mathrm{AM}} = [A_3^{\mathrm{HT}}, (\tilde{S}X_1, \tilde{S}X_2)^0]$, $\bar{S} = I_{NT} - S_2 - J$ and $\tilde{S} = I_{NT} - S_3 - J$. The order condition for γ becomes:

$$K + k_1 + k_3 + g_1 + (T-1)(k_1 + k_3) \geq K + G \text{ or } T(k_1 + k_3) \geq g_2,$$

and for parameters δ, we must have:

$$K + k_2 + h_1 + (N-1)(k_1 + k_2) \geq K + H \text{ or } N(k_1 + k_2) \geq h_2.$$

Now, and following BMS, if X_2, X_4 are correlated with the individual effect α only through a time-invariant component, and if (X_3, X_4) are correlated with the time effect λ only through a individual-invariant component, the BMS-like instruments are equivalent to the expanded instruments sets:

$$A^{\mathrm{BMS}} = (A_1^{\mathrm{BMS}}, A_2^{\mathrm{BMS}}, A_3^{\mathrm{BMS}}),$$

where $A_1^{\mathrm{BMS}} = (S_1 X) = A_1^{\mathrm{HT}} = A_1^{\mathrm{AM}}$, $A_2^{\mathrm{BMS}} = [A_2^{\mathrm{AM}}, (\bar{S}X_2, \bar{S}X_4)^*]$ and $A_3^{\mathrm{BMS}} = [A_3^{\mathrm{AM}}, (\tilde{S}X_3, \tilde{S}X_4)^0]$. The order condition for these instruments is $T(k_1 + k_3) + (\text{T-1})(k_2 + k_4) \geq g_2$ for γ and $N(k_1 + k_2) + (N-1)(k_3 + k_4) \geq h_2$ for δ.

In order to compute $\hat{v}_{\mathrm{IV}}$, we must first estimate parameters θ_1^2, θ_2^2 and θ_3^2. To do this, we can use a consistent estimate of β, δ and γ, and estimates of the variance components derived from these estimators will be used below for estimating $\hat{\varphi}_{\mathrm{IV}}$. We can summarize the complete procedure as follows:

- Compute the within estimator $\hat{\beta}_W = (X' S_1 X)^{-1}(X' S_1 Y)$ and form the vector of residuals $\hat{u}_w = S_1 Y - S_1 X \hat{\beta}_W$ to compute

$$\hat{\theta}_1^2 = \hat{\sigma}_\varepsilon^2 = (\hat{u}_w' \hat{u}_w)/(N-1)(T-1) - K. \tag{4.32}$$

- Regress $S_2 Y - S_2 X \hat{\beta}_W$ on $P_{A_2} Z$ to get a consistent estimate $\hat{\gamma}_{\mathrm{IV}}$ and form the residuals vector $\hat{u}_2 = S_2 Y - S_2 X \hat{\beta}_W - S_2 Z \hat{\gamma}_{\mathrm{IV}}$. We can show that for fixed T and $N \longrightarrow \infty$:

$$\operatorname{plim}(\hat{u}_2' \hat{u}_2 / N) = \theta_2^2.$$

- Regress $S_3 Y - S_3 X \hat{\beta}_W$ on $P_{A_3} Z$ to get a consistent estimate $\hat{\delta}_{IV}$ and form vector of residuals $\hat{u}_3 = S_3 Y - S_3 X \hat{\beta}_W - S_3 W \hat{\delta}_I$. We can show that for fixed N and $T \longrightarrow \infty$:

$$\operatorname{plim}(\hat{u}_3' \hat{u}_3 / T) = \theta_3^2.$$

4.6 GMM Estimation of Static Panel Data Models

The way to deal with correlated effects using an IV procedure is to construct orthogonality conditions from the model residual and instruments such as those presented above (HT, AM, BMS), those instruments being assumed uncorrelated

with the disturbances (at least asymptotically), and asymptotically correlated with explanatory variables. Consistent parameter estimates are then obtained under the assumption that the model is correctly specified (i.e., that orthogonality conditions are valid), by minimizing a quadratic form in orthogonality conditions (moment restrictions). Depending on the way this criterion is constructed, we obtain either either the Instrumental Variables (IV) under various forms, or the Generalized Method of Moments estimator (GMM, see Hansen (1982)). We now turn to the application of GMM estimation to linear panel data models.

4.6.1 Static Model Estimation

We consider here the general form of Instrumental Variable and GMM estimators for the static model introduced above, $Y = X\beta + Z\gamma + U$, or in a compact form, $Y = \Phi\mu + U$ where $\Phi = (X,Z)$ and $\mu' = (\beta', \gamma')$. Let $E(A_i'U_i) = 0$ denote a L set of orthogonality conditions in vector form, where A_i, $i = 1,2,\ldots,N$ is a $T \times L$ matrix of instruments. For a fixed T and $N \to \infty$, the empirical counterpart of the orthogonality conditions is $(1/N)\sum_{i=1}^{N} A_i'U_i$.

Consider estimating by Generalized Least Squares (GLS) the following equation:

$$A'Y = A'X\beta + A'Z\gamma + A'U = A'\Phi\mu + A'U, \tag{4.33}$$

i.e., by minimizing

$$\min \frac{1}{N}U'A\left[\mathrm{Var}\left(\frac{1}{N}A'U\right)\right]^{-1}\frac{1}{N}A'U. \tag{4.34}$$

Letting V denote the variance–covariance matrix of $(1/N)A'U$, the resulting estimator can be written as

$$\hat{\mu} = (\Phi'AV^{-1}A'\Phi)^{-1}\Phi'AV^{-1}A'Y$$

Suppose we do not wish to make assumptions on the structure of the variance matrix V, e.g., disturbances may exhibit serial correlation (in the time dimension) and/or heteroskedasticity. Then the estimator above can be computed using an initial estimate for V, $\hat{V} = (1/N)\sum_{i=1}^{N} A_i'\hat{U}_i\hat{U}_i'A_i$, where A_i is the (T,L) matrix of observations about the instrumental variables for the i-th individual and $\hat{U}_i$ is a $(T,1)$ initial consistent estimate of U_i, $i = 1,\ldots,N$. This estimator is the GMM (Generalized Method of Moments) under its optimal form, and its exploits the fact that the variance–covariance matrix of U_i is block-diagonal (no correlation across individuals).

It is well known that if the disturbances are both homoskedastic and not serially correlated, so that

$$\mathrm{Var}(A'U) = E[A'\mathrm{Var}(U|A)A] + \mathrm{Var}[A'E(U|A)A] = \sigma^2 E(A'A)^{-1}$$

since $E(U|A) = 0$ and $\text{Var}(U|A) = \sigma_U^2 I$, then the "best" instrumental variables estimator (i.e., GLS applied to model) is given by (see Gourieroux and Monfort, 1989):

$$\hat{\mu}_{\text{IV}} = \left[\Phi' A (A'A)^{-1} A' \Phi\right]^{-1} \Phi' A (A'A)^{-1} A' Y$$

For panel data however, this is unlikely to be the case because of individual unobserved heterogeneity, and the variance–covariance matrix of error terms is $\Omega = I_N \otimes \Sigma$ where $\Sigma = \sigma_\alpha^2 e_T e_T' + \sigma_\varepsilon^2 I_T$ is a $T \times T$ matrix. Suppose a preliminary estimate of Ω is available, $\hat{\Omega} = I_N \otimes \Sigma\left(\hat{\sigma}_\alpha^2, \hat{\sigma}_\varepsilon^2\right)$, a simple version of this being $\hat{\Omega} = I_N \otimes \frac{1}{N}\sum_{i=1}^{N} \hat{U}_i \hat{U}_i'$. Replacing Ω by $\hat{\Omega}$ so that

$$\text{plim}_{N\to\infty} \frac{1}{N} A' \Omega A = \text{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} A_i' \hat{\Omega} A_i = V,$$

we obtain the Three-Stage Least Squares estimator:

$$\hat{\mu}_{\text{3SLS}} = \left[\Phi' A \left(A' \Omega A\right)^{-1} A' \Phi\right]^{-1} \Phi' A \left(A \Omega A'\right)^{-1} A' Y\,.$$

It is easy to see that the GMM and the 3SLS are equivalent under the condition of *no conditional heteroskedasticity*, see Ahn and Schmidt (1999):

$$E\left(A_i' U_i U_i' A_i\right) = E\left(A_i' \Sigma A_i\right) \quad \forall i = 1, \ldots, N.$$

Note that this condition is weaker than the condition that $E(U_i U_i'|A_i) = \Sigma$. If the no-conditional heteroskedasticity condition is not satisfied, then GMM is more efficient than 3SLS.

Assuming this conditional holds, 2SLS estimators can also be proposed. A first version of the Two-Stage Least Squares (2SLS) estimator is obtained by premultiplying the model by $\Omega^{-1/2}$ and then applying instruments A. This is the form used by HT, AM and BMS:

$$\hat{\mu}_{\text{IV}_1} = \left[\Phi' \Omega^{-1/2} A (A'A)^{-1} A' \Omega^{-1/2} \Phi\right]^{-1} \Phi' \Omega^{-1/2} A (A'A)^{-1} A' \Omega^{-1/2} Y$$

This estimator is based on the two conditions:

$$E\left(A_i' \Sigma^{-1/2} U_i\right) = 0, \qquad E\left(A_i' \Sigma^{-1/2} U_i U_i' \Sigma^{-1/2} A_i\right).$$

Ahn and Schmidt (1999) shows that the 3SLS and the 2SLS estimator above are equivalent asymptotically if a consistent estimate is used for Σ[4] and if there exists a nonsingular and non-random matrix B such that $\Omega^{-1/2} A = AB$ (or equivalently, that $\Sigma^{-1/2} A_i = A_i B \quad \forall i = 1, 2, \ldots, N$).

A second version of the 2SLS estimator is denoted Generalized Instrumental Variables (GIV, see White (1984)), which uses directly $\Omega^{-1/2} A$ as instruments:

[4] They are numerically equivalent if the same, consistent estimate is used.

$$\hat{\mu}_{\mathrm{GIV}} = \left[\Phi'^{-1}A\left(A'^{-1}A\right)^{-1}A'^{-1}\Phi\right]^{-1}\Phi'^{-1}A\left(A'^{-1}A\right)^{-1}A'^{-1}Y.$$

Although the two 2SLS estimators seem different, they are equivalent in a panel-data context when the error component structure is of the form Ω above. Again, a preliminary estimate of Σ is required to implement these 2SLS estimation procedures.

4.6.2 GMM Estimation with HT, AM and BMS Instruments

In the Instrumental-Variable context with Hausman–Taylor, Amemiya–MaCurdy or Breusch–Mizon–Schmidt instruments described above, we assume an error-component structure and also that endogeneity is caused by correlated effects, either $E(X'\alpha) \neq 0$ or $E(Z'\alpha) \neq 0$. In any case, it is maintained that $E(X'\varepsilon) = E(Z'\varepsilon) = 0$. With GMM, we can consider different exogeneity assumptions related to α or ε, producing different orthogonality conditions. Apart from the difference between random and fixed effect specifications (instruments correlated or not with α), we can also consider strictly or weakly exogenous instruments if explanatory instruments are correlated with ε. These different cases are described by Ahn and Schmidt (1999), to which we refer the reader for more information.

Consider the case of strict exogeneity: $E(X_{is}\varepsilon_{it}) = E(Z_i\varepsilon_{it}) = 0$, $\forall i, \forall t$. The questions we address are the following: is it possible to obtain a more efficient estimator than IV with either HT, AM or BMS, by exploiting more moment conditions? And does this efficiency depend on the assumption made on the assumed variance–covariance structure?

The first result is that, under the *No conditional heteroskedasticity* assumption, HT, AM and BMS–2SLS estimators are equivalent to the GMM estimator. From the discussion above, this implies that GMM is more efficient with the same instrument set (and a consistent variance–covariance matrix) than the original version of HT, AM and BMS–2SLS estimators, if this NCH condition is not valid.

Ahn and Schmidt (1995) and Arellano and Bover (1995) note that, under the strict exogeneity assumption, more moment conditions can be used, to improve efficiency of the estimator. The strict exogeneity assumption is $E(d_i \otimes \varepsilon_i) = 0$, where $d_i = (x_{i1}, \ldots, x_{iT}, z_i)$, implying $E[(L_T \otimes d_i)'u_i] = E(L_T'\varepsilon_i \otimes d_i) = 0$.

Arellano and Bover (1995) therefore propose a GMM estimator obtained by replacing (in vector form) $Q_T\Phi_i$ by $L_T \otimes d_i$ in the HT, AM or BMS list of instruments. This leads to $(T-1)(kT+g)-k$ additional instruments, which may cause computational difficulties if T is large. They however also show that under the error-component structure Σ, both sets of instruments provide the same 3SLS (or 2SLS version 1, $\hat{\mu}_{\mathrm{IV1}}$) estimator.[5] Consequently, if in addition the *No conditional heteroskedasticity* assumption is valid, then the 3SLS (or 2SLS version 1, $\hat{\mu}_{\mathrm{IV1}}$) with HT, AM or BMS instruments will be asymptotically equivalent to GMM with the augmented set of instruments.

[5] Asymptotically only, if different estimates of Σ are used.

Im, Ahn, Schmidt and Wooldridge (1999) consider cases where the *no conditional heteroskedasticity* assumption holds when the Arellano–Bover set of instruments is used, and Σ is left unrestricted. They show that the 2SLS estimator version 2 ($\hat{\mu}_{\mathrm{IV2}}$) using BMS instruments is equivalent to the 3SLS estimator using Arellano–Bover instruments, but that this equivalence does not hold for HT or AM instruments. To solve this problem, Im, Ahn, Schmidt and Wooldridge (1999) propose to replace the fixed effects operator Q_T by $Q_\Sigma = \Sigma^{-1} - \Sigma^{-1} e_T (e_T' \Sigma^{-1} e_T)^{-1} e_T' \Sigma^{-1}$, such that $Q_\Sigma e_T = 0$, and modifying the matrix of instruments appropriately. This modified 3SLS estimator would be asymptotically equivalent to an efficient GMM estimation is the NCH condition holds.

4.7 Unbalanced Panels

In the preceding sections we have discussed estimation methods for panel data models when all cross-sectional units are observed for all time periods. In practice, missing observations are often encountered for a given cross-sectional unit and for a given time period. In this case, we have what we call an incomplete panel and the standard estimation methods are not applicable. Fuller and Battese (1974) suggest to add in the list of regressors a set of dummy variables, one for each missing observation. However, as noted by Wansbeek and Kapteyn (1989), this often implies that the number of regressors would increase dramatically (possibly, with the sample size), and in many empirical studies this becomes computationally impractical.

Wansbeek and Kapteyn (1989) consider a two-way unbalanced error component specification for the fixed and random effects models. In the first case (fixed effects) they suggest a new expression for within operator, which generalizes the operator Q given in Sect. 4.2.[6] For the second case (random effects), they propose to use the quadratic unbiased and Maximum Likelihood estimators.

More recently, Baltagi and Chang (1994) have considered a one-way error component model with unbalanced data. Using a Monte Carlo simulation experiment, they compare several estimation methods including the Analysis Of Variance (ANOVA), Maximum Likelihood (ML), Restricted Maximum Likelihood (REML), Minimum Norm Quadratic Estimation (MINQUE) and Minimum Variance Quadratic Estimation (MINQUE).

In their simulation and the empirical illustration they propose, they show that in general, better estimates of the variance components do not necessarily imply better estimates of the regression coefficients. Furthermore, MLE and MIVQUE perform better than the ANOVA methods in the estimation of the individual variance component. Finally, for the regression coefficients, the computationally simple ANOVA methods perform reasonably well when compared with the computationally involved MLE and MIVQUE methods.

When the data have a sufficient degree of disaggregation, more than two dimensions of data variation are generally available. One can think for instance of a sample

[6] See Wansbeek and Kapteyn (1989), p. 344.

of observations on firms (level 1) belonging to a particular industry (level 2), within a region (level 3). In this case, several time-invariant heterogeneity components can be introduced in the linear panel data model, giving rise to multi-way error components models. In the nested specification, each successive component in the error term is nested within the preceding component. In the non-nested case, error components are independent of each other, and transformation techniques similar to those employed in the two-way error component model are applicable.

As operator matrices for performing Between and Within transformations under any hierarchical structure are straightforward to construct, fixed effects and GLS estimators are generally available for such models (see, e.g., Antweiler (2001)). In the unbalanced panel data case however, the required algebra to obtain expressions for the Feasible GLS estimator in particular, is more difficult to handle.

Baltagi, Song and Jung (2001) propose a fixed-effects representation and a spectral decomposition of a three-way unbalanced error component model, leading to a Fuller–Battese scalar transformation for this model. They proceed by investigating the performance of ANOVA, Maximum Likelihood and MINQUE estimators of variance components in the unbalanced nested error component model. ANOVA estimators for variance components are BQU (Best Quadratic Unbiased) in the balanced case only, and are only unbiased in the unbalanced case. Monte Carlo experiments reveal that ANOVA methods perform well in estimating regression coefficients, but ML and MINQUE estimators are recommended for variance components and standard errors of regression coefficients. They do not deal with the case of endogenous regressors or correlated effects, beyond the obvious possibility to obtain consistent estimates using fixed effects. The fact that exogenous variables may be available for different levels in the hierarchical structure of the data, leads to a wide variety of possible instruments. For example, if firm-specific individual effects are correlated with decision variables of the firms, price variables at an upper level (county, region) may be used as instruments.

Davis (2002) proposes a unifying approach to estimation of unbalanced multi-way error components models, as well as useful matrix algebra results for constructing (Between, Within) transformation matrices. The recurrence relations proposed in the paper allow for direct extension to any number of error components. There are but few empirical applications in the literature using multi-way unbalanced panels, see Davis (2002) and Boumahdi and Thomas (2006) for examples.

For example, the three-way unbalanced error component model is

$$Y = X\beta + u, \qquad u = \Delta_1\alpha + \Delta_2\gamma + \Delta_3\lambda + \varepsilon, \tag{4.35}$$

where $\alpha = (\alpha_1, \ldots, \alpha_L)'$, $\gamma = (\gamma_1, \ldots, \gamma_H)'$ and $\lambda = (\lambda_1, \ldots, \lambda_T)'$.

Matrices $\Delta_k, k = 1,2,3$ are constructed by collecting dummy variables for the relevance of a given observation to a particular group (l,h,t), and have dimension $N \times L$, $N \times H$ and $N \times T$ respectively.

Letting $P_A = A(A'A)^+A'$ and $Q_A = I - P_A$ where $^+$ denotes a generalized inverse, the fixed effects transformation matrix is shown to be $Q_\Delta = Q_A - P_B - P_C$, where $\Delta = [\Delta_1, \Delta_2, \Delta_3]$ and

$$P_A = I - \Delta_3(\Delta_3'\Delta_3)^{+}\Delta_3', \quad Q_A = I - P_A,$$
$$P_B = Q_A\Delta_2(\Delta_2'Q_A\Delta_2)^{+}\Delta_2'Q_A, \quad Q_B = I - P_B,$$
$$P_C = Q_AQ_B\Delta_1\left[\Delta_1'(Q_AQ_B)\Delta_1\right]^{+}\Delta_1'Q_AQ_B, \quad Q_C = I - P_C.$$

Under the exogeneity assumption $E(X'Q_\Delta\varepsilon) = 0$, the fixed-effects estimator is consistent:

$$\hat{\beta} = \left(X'Q_\Delta X\right)^{-1}X'Q_\Delta Y. \tag{4.36}$$

Assume instruments W are available such that $E(W'Q_\Delta\varepsilon) = 0$; then a consistent IV estimator can be constructed as

$$\hat{\beta} = \left(X'P_{QW}X\right)^{-1}X'P_{QW}Y, \tag{4.37}$$

where $P_{QW} = Q_\Delta W\left(W'Q_\Delta W\right)^{-1}W'Q_\Delta$.

As mentioned above in the one-way unbalanced case, application of IV procedures require consistent estimation of the variance–covariance matrix, as well as an instrument matrix consistent with the unbalanced nature of the sample. Formulae for estimating variance components can be found in Baltagi, Song and Jung (2001) and Davis (2002), although estimation should be adapted along the lines of Hausman and Taylor (1981) because of endogenous regressors. For instrument matrices, the HT specification is directly applicable because it only contains Within transformations and variables in levels. However, the AM and BMS IV estimators suffer from the same difficulty as in the one-way unbalanced case: they are more problematic to adapt because of missing observations in X_1^* and $(QX)^*$ matrices. It is not clear whether the usual procedure to replace missing values by zeroes in those matrices produces severe distortions (bias, inefficiency) or not.

References

Ahn, S.C., Y.H. Lee and P. Schmidt [2001]. GMM Estimation of Linear Panel Data Models with Time-Varying Individual Effects. Journal of Econometrics 101, 219–255.

Ahn, S.C. and P. Schmidt [1999]. Estimation of linear panel data models using GMM. In Generalized Method of Moments Estimation (L. Matyas, ed.), pp. 211–247, Cambridge University Press, Cambridge.

Alvarez, J. and M. Arellano [2003]. The Time Series and Cross Section Asymptotics of Dynamic Panel Data Estimators. Econometrica 71, 1121–1159.

Amemiya, T. and T.E. MaCurdy [1986]. Instrumental Variable Estimation of an Error Components Model. Econometrica 54, 869–880.

Anderson, T. and C. Hsiao [1982]. Formulation and Estimation of Dynamic Models Using Panel Data. Journal of Econometrics 18, 47–82.

Antweiler, W. [2001]. Nested Random Effects Estimation in Unbalanced Panel Data. Journal of Econometrics 101, 295–313.

Arellano, M. [1993]. On the Testing of Correlated Effects with Panel Data. Journal of Econometrics 59, 87–97.

Arellano, M. and S. Bond [1991]. Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. Review of Economic Studies 58, 277–297.

Arellano, M. and O. Bover [1995]. Another Look at the Instrumental Variable Estimation of Error-Components Models. Journal of Econometrics 68, 29–51.

Baltagi, B.H. and Y.-J. Chang [1994]. A Comparative Study of Alternative Estimators for the Unbalanced One-Way Error Component Regression Model. Journal of Econometrics 62, 67–89.

Baltagi, B. and S. Khanti-Akom [1990]. On Efficient Estimation with Panel Data, an Empirical Comparison of Instrumental Variables Estimators. Journal of Applied Econometrics 5, 401–406.

Baltagi, B.H., S.H. Song and B.C. Jung [2001]. The Unbalanced Nested Error Component Regression Model. Journal of Econometrics 101, 357–381.

Biørn, E. [1981]. Estimating Economic Relations from Incomplete Cross-Section/Time-Series Data. Journal of Econometrics 16, 221–236.

Boumahdi, R., J. Chaaban and A. Thomas [2006]. Import demand estimation with country and product effects: Application of multi-way unbalanced panel data models to Lebanese imports. In Panel Data Econometrics: Theoretical Contributions and Empirical Applications (B.H. Baltagi, ed.), Chap. 8, pp. 193–228, Elsevier, Amsterdam.

Boumahdi, R. and A. Thomas [1997]. Estimation des modèles de données de panel avec régresseurs temporels. Annales d'Economie et de Statistique 46, 23–48.

Boumahdi, R. and A. Thomas [2006]. Instrument Relevance and Efficient Estimation with Panel Data. Economics Letters 93, 305–310.

Breusch, T.S., G.E. Mizon and P. Schmidt [1989]. Efficient Estimation Using Panel Data. Econometrica 57, 695–700.

Cornwell C. and P. Rupert [1988]. Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variables Estimators. Journal of Applied Econometrics 3, 149–155.

Crépon, B., F. Kramarz and A. Trognon [1997]. Parameters of Interest, Nuisance Parameters and Orthogonality Conditions an Application to Autoregressive Error Component Models. Journal of Econometrics 82, 135–156.

Davis, P. [2002]. Estimating Multi-Way Error Components Models with Unbalanced Data Structures. Journal of Econometrics 106, 67–95.

Fuller, W.A. and G.E. Battese [1974]. Estimation of Linear Functions with Cross-Error Structure. Journal of Econometrics 2, 67–78.

Godfrey, L.G. [1999]. Instrument Relevance in Multivariate Linear Models. Review of Economics and Statistics 81, 550–552.

Hansen, L.P. [1982]. Large Sample Properties of Generalized Method of Moments Estimators. Econometrica 50, 1029–1054.

Hausman, J. [1978]. Specification Tests in Econometrics. Econometrica 46(6), 1251–1271.

Hausman, J. and W.E. Taylor [1981]. Panel Data and Unobservable Individual Effects. Econometrica 49, 1377–1398.

Heckman, J.J. and V.J. Holtz [1989]. Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. Journal of the American Statistical Association 84, 862–875.

Holtz-Eakin, D., W. Newey and H. Rosen [1988]. Estimating Vector Autoregressions with Panel Data. Econometrica 56, 1371–1395.

Hsiao, C. [1986]. *Analysis of Panel Data*. Cambridge University Press, Cambridge.

Im, K.S., S.C. Ahn, P. Schmidt and J.M. Wooldridge [1999]. Efficient Estimation of Panel Data Models with Strictly Exogenous Explanatory Variables. Journal of Econometrics 93, 177–201.

Lillard, L. and Y. Weiss [1979]. Components of Variation in Panel Earnings Data: American Scientists 1960–1970. Econometrica 47, 437–454.

Mundlak, Y. [1961]. Empirical Production Function Free of Management Bias. Journal of Farm Economics 43, 44–56.

Mundlak, Y. [1978]. On the Pooling of Time Series and Cross Section Data. Econometrica 46, 69–85.

Nauges, C. and A. Thomas [2003]. Consistent Estimation of Dynamic Panel Data Models with Time-Varying Individual Effects. Annales d'Economie et de Statistique 70, 53–75.

Schmidt, P., S.C. Ahn and D. Wyhowski [1992]. Comment: Sequential Moment Restrictions in Panel Data. Journal of Business and Economic Statistics 10, 10–14.
Shea, J. [1997]. Instrument Relevance in Multivariate Linear Models: A Simple Measure. Review of Economics and Statistics 79, 348–352.
Wansbeek, T. and A. Kapteyn [1989]. Estimation of the Error-Components Model with Incomplete Panels. Journal of Econometrics 41, 341–361.
Wansbeek, T.J. and T. Knaap [1999]. Estimating a Dynamic Panel Data Model with Heterogenous Trends. Annales d'Economie et de Statistique 55–56, 331–349.
Wyhowski, D.J. [1994]. Estimation of a Panel Data Model in the Presence of Correlation Between Regressors and a Two-Way Error Component. Econometric Theory, 10, 130–139.

Chapter 5
The Chamberlain Approach to Panel Data: An Overview and Some Simulations

Bruno Crépon and Jacques Mairesse

5.1 Introduction

In this paper, we present the general approach proposed by Chamberlain (1982 and 1984) for the analysis of panel data. Although the 1984 article examines nonlinear models such as probit or logit models for panel data, we only cover here, but in details, the case of linear regression models. However, much of the approach which is applicable in this case, as well as much of the intuition that can be gained from it, applies also to non linear models.

Let us consider the linear regression model of the dependent variable y on K explanatory variables x observed for a balanced panel of N individuals (for example firms) and T periods (for example years):

$$y_{it} = \sum_k x_{it}^{(k)} b_k + v_{it} = \sum_k x_{it}^{(k)} b_k + \mu_i + u_{it}, \quad t = 1, \ldots, T, \quad i = 1, \ldots, N \qquad (5.1)$$

The Chamberlain approach or method provides a general unifying framework encompassing both the ideal case of reference of "Non Correlated Errors" (NCE) in which the regressors x in model (5.1) can be assumed to be uncorrelated with both the unobserved individual effects μ and the idiosyncratic disturbances u (i.e., with the overall disturbance v), and the paradigmatic case in panel data econometrics of "Correlated Effects" (CE) in which the regressors x are possibly correlated with the individual effects μ but are still assumed to be uncorrelated with the idiosyncratic disturbances u (see Mundlak, 1961). It allows to take into account different important types of specification errors other than the existence of correlated effects, in the form of correlations or covariances between the regressors x and both disturbances

Bruno Crépon
CREST-INSEE (Paris), CEPR (London) and IZA (Bonn), e-mail: crepon@ensae.fr

Jacques Mairesse
CREST-INSEE (Paris), UNU-MERIT (Maastricht University), and NBER. CREST-INSEE, 15 boulevard Gabried PERI, 92245 MALAKOFF cédex, France, e-mail: mairesse@ensae.fr

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

μ and u (or v). The variables x, for example, can be affected by random errors of measurement, or they can include lags of the dependent variable y.

The Chamberlain method consists in a first stage in constructing the set (or a subset) of relations between the moments of the variables x and the unknown parameters b, which are implied by the correlations or covariances between the regressors and the disturbances. In a second stage it uses these relations, or estimating equations, to implement Minimum Distance estimators, or so called Asymptotic Least Squares (ALS) estimators, and obtain consistent and possibly asymptotically efficient estimates of the b's. The Chamberlain method is thus basically a method of moments; and it is in fact very close to the generalized methods of moments (GMM) as applied to panel data, for which it provides a different, deeper understanding and a possible alternative.

In Sect. 5.2, we present the first stage of the Chamberlain method, consisting itself of two steps. We first define the so called Chamberlain's matrix $\Pi = E(\underline{y}_i\underline{x}_i')E(\underline{x}_i\underline{x}_i')^{-1}$ of dimension (T, KT), providing a summary of the panel data.[1] We then derive the Chamberlain estimating equations, which are the basic relations existing between the parameters b of (primary) interest to be estimated and the coefficients of Π, or equivalently, and often more simply, the relations between the b' s and the coefficients of $E(\underline{y}_i\underline{x}_i')$ and $E(\underline{x}_i\underline{x}_i')$, that is the covariance matrices of y and the x's across all T periods. We explicitly consider these estimating equations in specific cases of major interest, corresponding to different types of specification errors imposing restrictions on the form of $E(\underline{v}_i\underline{x}_i')$,the covariance matrix of the disturbance v and the x's across all T periods. In this section, we also explain how additional restrictions arising from specific assumptions on the form of the covariance matrix $E[\underline{v}_i\underline{v}_i']$ of the disturbances can be used to improve the efficiency of the Chamberlain method estimators. We also provide an extended view of the Chamberlain method which applies to more general models than the regression model (5.1).

In Sect. 5.3, we present the second stage of the Chamberlain method, consisting in applying the Asymptotic Least Squares (ALS) to obtain consistent and asymptotically normal estimators of the parameters b's of primary interest, as well as parameters of secondary interest characterizing possible specification errors in model (5.1). We explain how to implement the asymptotically efficient or optimal ALS estimator and perform specification tests. We also explain how the ALS estimating equations can be manipulated and some parameters eliminated without loosing consistency and asymptotic efficiency in estimating the remaining parameters.

In Sect. 5.4, we show how in particular the ALS estimating equations can be reformulated as orthogonality conditions in the panel data GMM framework, and we demonstrate explicitly the asymptotic equivalence between the Chamberlain method and GMM in general and in some of the specific cases of interest.

In the last Sect. 5.5, we present some simulation results illustrating the small sample properties of both the Chamberlain method and the GMM estimators in some of the specific cases of interest previously considered.

---

[1] $\underline{y}_i$, $\underline{v}_i$ and $\underline{x}_i$ are column vectors of dimension $(T, 1)$ and dimension $(KT, 1)$ respectively. See Sects. 5.2.1 and 5.2.2 below.

The Appendices A, B, C, D and E (respectively Sect. 5.6 to 5.10) provide details on various technical points. They can be skipped by readers who only want to have an overall understanding of the Chamberlain method and its equivalence with GMM. In Appendix A, we show how the Chamberlain approach can be extended to simultaneous linear equations models and to vector autoregression models (VAR) on panel data, and how it can deal also with endogenous attrition. In Appendix B, we show how the Chamberlain estimating equations, written in matrix format in the text, can also be written in a vector format for practical estimation. In Appendix C, we show how the Chamberlain estimating equations can be rewritten in order to eliminate auxiliary parameters (or parameters of secondary interest), while preserving asymptotic consistency and efficiency of the estimators of the parameters of (primary) interest. In Appendix D, we show that the usual basic panel estimators (Random Effects, Within and First Differences) are asymptotically equivalent to the Chamberlain and GMM estimators. In Appendix E, we provide important details on the design and calibration of the simulation experiments presented in the text (Sect. 5.5).

5.2 The Chamberlain Π Matrix Framework

5.2.1 The Π Matrix

The Chamberlain method is basically a method of moments; it uses the restrictions on the moments of the variables implied by the modeling assumptions to estimate the parameters of interest. In a first stage, the moments of the variables are computed up to the second order forming a set of summary statistics for the data, which can be considered as estimated auxiliary parameters. In a second stage the parameters of interest are estimated on the basis of their relations with these estimated auxiliary parameters, using Minimum Distance or Asymptotic Least Squares (ALS) estimators.

An important feature of the Chamberlain method, as usually presented, is that it summarizes the set of second order moments (variances and covariances) of the dependent and explanatory variables, which is central to the analysis, by the so called Chamberlain Π matrix. The Π matrix is defined in terms of the coefficients of the linear predictors of the dependent variable at each period given all explanatory variables at all periods. Precisely, if there are T years and K explanatory variables x, the $\Pi = [\pi_{t,j}]$ matrix is of dimension $T \times KT$ and is obtained by stacking one above the other the row vectors of dimension $1 \times KT$ of the coefficients of the T separate year regressions such as

$$y_{it} = \pi_{t,1} x_{i1}^{(1)} + \cdots + \pi_{t,T} x_{iT}^{(1)} + \pi_{t,T+1} x_{i1}^{(2)} + \cdots + \pi_{t,KT} x_{iT}^{(K)} + w_{it} \tag{5.2}$$

with

$$E(w_{it} x_{is}^{(k)}) = 0, \quad \forall s,t,k. \tag{5.3}$$

If we define $\underline{y}_i' = (y_{i1}, \ldots, y_{iT})$ and $\underline{x}_i' = (x_{i1}^{(1)}, \ldots, x_{iT}^{(1)}, \ldots, x_{iT}^{(K)})$ we can also write

$$\Pi = E(\underline{y}_i \underline{x}_i') E(\underline{x}_i \underline{x}_i')^{-1}. \tag{5.4}$$

It must be noted that (5.2) with the covariance restriction (5.3) is not a linear regression model strictly speaking, but simply expresses the linear projection of y on all lagged, present and future x 's.[2]

As a simple illustration, let us take the example of a regression model for a panel of only two years (T=2), with only one explanatory variable (K=1) and the usual error components structure

$$y_{it} = x_{it} b + \mu_i + u_{it}, \quad t = 1,2, \quad i = 1, \ldots, N.$$

Consider first the simplest standard case that we call "Non Correlated Errors" (NCE), in which it is assumed that

$$E[\mu_i x_{is}] = E[u_{it} x_{is}] = 0, \quad \forall t, s.$$

In this case, the parameter of interest b can be consistently estimated by simply using the pooled OLS estimator or the error components GLS estimator. However, it is also possible to compute the Π matrix by considering the two separate year regressions with both x_1 and x_2 as explanatory variables. In each year regression the true coefficient of the contemporaneous x (π_{11} or π_{22}) is equal to the parameter b and the other coefficient of the lagged or future x (π_{21} or π_{12}) is zero. Thus the "true" Π matrix is

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} = \begin{bmatrix} b & 0 \\ 0 & b \end{bmatrix} = b \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

We thus can expect that the unrestricted Π matrix, as estimated in the first stage, will also look roughly diagonal (and with roughly equal diagonal coefficients). However, this may not be so striking in practice, if the individual $\pi_{t,s}$ coefficients are not estimated precisely enough, that is if the sample is not large enough in the individual dimension and if the year x's (x_1 and x_2) are more or less collinear (which is often likely when these variables are in absolute level or defined as ratio of different variables). In the second stage of the Chamberlain method, we can retrieve a consistent estimate $\widehat{b}$ of b from the estimated $\widehat{\pi}_{t,j}$ by applying Asymptotic Least Squares (ALS) to the four "estimating equations" $\pi_{11} = b$, $\pi_{12} = 0$, $\pi_{21} = 0$, $\pi_{22} = b$, or in vector form

$$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})' = b(1,0,0,1)'.$$

The ALS estimator is defined precisely in the next Sect. 5.3. It is more efficient than the pooled OLS estimator, and also more efficient than the usual error components GLS estimator, under general conditions in which v_{it} has not an error

[2] The linear projection is generally denoted by $E^*(y_{it} | x_{i\tau}^{(k)}; \tau = 1, \ldots, T; k = 1, \ldots, K)$ with a * to distinguish it from the conditional expectation $E(y_{it} | x_{i\tau}^{(k)}; \tau = 1, \ldots, T; k = 1, \ldots, K)$, which has no a priori reason to be linear in the x's. We do not, however, use this notation here.

component structure and is heteroscedastic and serially correlated.[3] One can also test formally whether the Π matrix has the right pattern, by an asymptotic χ^2 test "measuring" the distance between the unrestricted first stage estimate of Π and its restricted second stage estimate.[4]

Let us consider next the other standard case of "Correlated Effects" (CE) in which the explanatory variable is correlated with the individual effects μ_i but not with the error terms u_{it}

$$E[u_{it}x_{is}] = 0, \quad \forall t,s \qquad E[\mu_i x_{is}] \neq 0.$$

In this case, the linear projection of the individual effects μ_i on the x's is no longer zero. It can be written as

$$\mu_i = \delta_1 x_{i1} + \delta_2 x_{i2} + \widetilde{w}_i$$

with, by definition, $E(\widetilde{w}_i x_{it}) = 0$.

Hence, in this case, the "true" Π matrix has the following distinctive pattern:

$$\begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} = \begin{bmatrix} \delta_1 + b & \delta_2 \\ \delta_1 & \delta_2 + b \end{bmatrix} = \begin{bmatrix} \delta_1 & \delta_2 \\ \delta_1 & \delta_2 \end{bmatrix} + \begin{bmatrix} b & 0 \\ 0 & b \end{bmatrix}$$

where the off-diagonal coefficients will have to be equal within the same columns for a panel with more than two years ($T > 2$).[5] As in the NCE case, such a distinctive pattern may be recognizable on the Π matrix as estimated in the first stage, although in general practice this pattern will be badly blurred.

Applying ALS to the four estimating equations

$$(\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})' = b(1,0,0,1)' + \delta_1(1,0,1,0)' + \delta_2(0,1,0,1)'$$

provides a consistent estimator of b, our parameter of primary interest, as well as a consistent estimator of the δ's. The coefficients δ's are usually called "nuisance parameters", but we will prefer to consider them here as parameters of secondary interest, since they characterize the relation of the unknown individual effects and the known explanatory variables. The ALS estimator is more efficient under general conditions than the usual Within estimator performed on the deviations of the variables from their individual means, that is by simply using pooled OLS on the within transformed regression of $y_{it} - y_{i.}$ on $x_{it} - x_{i.}$ (see Appendix D: Equivalence between Chamberlain's, GMM and usual panel data estimators).

[3] The usual error components GLS estimator is optimal under the assumption that μ_i and u_{it} are homoscedastic and u_{it} is serially uncorrelated.

[4] Note also that we may allow b to vary over time and test for its constancy.

[5] Note that the model with Correlated Effects is identified as soon as we have a panel: $T > 1$, and even if we allow the $\underline{b}$ coefficients to vary over time.

5.2.2 *Relations Between* Π *and the Parameters of Interest*

In this subsection, we generalize the two previous examples. We present the basic relations which the model implies between the second order moments of the variables y and x [i.e., $E(\underline{y}_i\underline{x}_i')$ and $E(\underline{x}_i\underline{x}_i')$] and the parameters of both primary interest and secondary interest, and we show how these relations can be rewritten as restrictions on the Π matrix.

Consider the linear regression model with K explanatory variables for a balanced panel of N individuals observed on T periods or years:

$$y_{it} = \sum_k x_{it}^{(k)} b_k + v_{it} = \sum_k x_{it}^{(k)} b_k + \mu_i + u_{it}, \quad t = 1,\ldots,T, \quad i = 1,\ldots,N \tag{5.5}$$

where the b's are the parameters of primary interest. A central idea of the Chamberlain approach is to view this panel data model as a system of T stacked year equations.

Defining the two $(T \times 1)$ column vectors $\underline{y}_i$ and $\underline{v}_i$ by $\underline{y}_i' = (y_{i1},\ldots,y_{iT})$ and $\underline{v}_i' = (\mu_i + u_{i1},\ldots,\mu_i + u_{iT})$ respectively, and the two $(KT \times 1)$ and $(K \times 1)$ column vectors by $\underline{x}_i' = (x_{i1}^{(1)},\ldots,x_{iT}^{(1)},\ldots,x_{i1}^{(1)},\ldots,x_{iT}^{(K)})$ and $\underline{b}' = (b_1,\ldots,b_K)$ respectively, we can more compactly rewrite the (5.5) as

$$\underline{y}_i = M(\underline{b})\underline{x}_i + \underline{v}_i \tag{5.6}$$

where $M(\underline{b}) = (b_1,\ldots,b_K) \otimes I_T$ is a $(T \times KT)$ matrix. Also denoting by $\Phi = E[\underline{v}_i\underline{x}_i']$ the $(T \times KT)$ covariance matrix between the disturbances $\underline{v}_i$ and explanatory variables $\underline{x}_i$, we can derive the following moments relations

$$E\left(\underline{y}_i\underline{x}_i'\right) = M(\underline{b})E\left(\underline{x}_i\underline{x}_i'\right) + \Phi. \tag{5.7}$$

Introducing now the matrix $\Pi = E(\underline{y}_i\underline{x}_i')E(\underline{x}_i\underline{x}_i')^{-1}$ and denoting by Ψ the (T×KT) matrix $E(\underline{v}_i\underline{x}_i')E(\underline{x}_i\underline{x}_i')^{-1} = \Phi E(\underline{x}_i\underline{x}_i')^{-1}$ of the coefficients of the linear projection of the disturbances $\underline{v}_i$ on the explanatory variables $\underline{x}_i$, we can also write equivalently

$$\Pi = M(\underline{b}) + \Psi. \tag{5.8}$$

The equations (5.5) or (5.6), or the moments relations (5.7) or (5.8), simply express an arbitrary decomposition of y into one part corresponding to the explanatory variables x and another one to the disturbances v. Giving them an econometric content (i.e., being able to test them as a regression model and to identify and estimate the b parameters of primary interest) requires imposing restrictions between the v's and the x's. In the Chamberlain method these stochastic restrictions can take different forms in terms of the Φ or Ψ matrices. The simplest case is the basic one of Non Correlated Errors (NCE) which assumes that $\Phi = \Psi = 0$.

More generally, let us consider that the $(T \times KT)$ covariance matrix Φ of $\underline{v}_i$ and $\underline{x}_i$ can be parameterized by a set of (additional) parameters $\underline{\beta}$ of secondary interest.[6] As long as the dimension of $\underline{\beta}$ is not too large (less than $KT^2 - K$), this implies the following set of restrictions between the moments of the variables $\underline{y}_i$ and $\underline{x}_i$

$$E(\underline{y}_i \underline{x}_i') = M(\underline{b}) E(\underline{x}_i \underline{x}_i') + \Phi(\underline{\beta}), \tag{5.9}$$

which can be rewritten as

$$\Pi = M(\underline{b}) + \Phi(\underline{\beta}) E(\underline{x}_i \underline{x}_i')^{-1}. \tag{5.10}$$

The core of the Chamberlain method is to derive estimates of the $\underline{b}$ and $\underline{\beta}$ parameters from the estimates of the Π matrix by applying Asymptotic Least Squares (ALS) to (5.10).[7] In the case, however, in which the matrix Ψ can be parameterized more conveniently in terms of parameters of secondary interest $\underline{\delta}$ than the matrix Φ in terms of parameters of secondary interest $\underline{\beta}$, estimates of the $\underline{b}$ and $\underline{\delta}$ parameters can be obtained from the estimated Π matrix alone by applying ALS to the equations

$$\Pi = M(\underline{b}) + \Psi(\underline{\delta}). \tag{5.11}$$

In this case, which is that of Correlated Effects (CE), the relations (5.11) take the form of a direct restriction on Π, and Π summarizes all the relevant information brought by the second order moments of the variables, with no need for the estimates of $E(\underline{x}_i \underline{x}_i')$.

In general, we have to rely on estimating equations such as (5.9) or (5.10) and to use estimates of $E(\underline{x}_i \underline{x}_i')$ or those of its inverse, which implies some additional complications compared with the more simple implementation of the ALS when the estimating equations can take the form of a direct restriction on Π as in (5.11). These complications are sometimes neglected or ignored, as we shall indicate in Sect. 5.3. Note that when Φ can be conveniently parameterized, usually Ψ cannot (and vice versa). The Correlated Effects case, which we are going to consider in more detail in the next subsection, is an exception where Φ and Ψ have the same structure where all T coefficients in a given column are equal to the corresponding β or δ coefficients. Note that in order to be able to identify fully the model (5.5) as specified, it is necessary to have at least one subset of the estimating equations (5.11) or (5.10) that can be solved for (b, δ) or (b, β) as a function of Π alone, or Π and $E(x_i x_i')$. A necessary (but not sufficient) condition for identification is that the dimension of the estimated parameter is less than the number of estimating

[6] A $(T \times KT)$ matrix can always be parameterized by its KT^2 coefficients. What we mean here is that it can be parameterized more parsimoniously.

[7] Note that it is equivalent to take (5.9) instead of (5.10) as estimating equations, with the only difference of using the estimated covariance matrix of y and x rather than the estimated Π matrix. The Π matrix has the advantage of being directly expressed in the set up of the regression model under analysis, with coefficients having the dimension of the regression coefficients (and not of covariances). But as we shall see it may be more convenient to rely on equations (5.9).

equations. Thus, as the number of equations (5.10) or (5.11) is KT^2 and the size of parameter $\underline{b}$ is K, the dimension of $\underline{\beta}$ or $\underline{\delta}$ must be less than $KT^2 - K$.[8]

The parameterization of the $T \times KT$ matrix $\Phi(\underline{\beta})$ or $\Psi(\underline{\delta})$ in the estimating equations (5.10) and (5.11) is essential in the implementation of the Chamberlain method. It expresses specific assumptions on the econometric specification of the model, some of which can be tested by the econometrician, while the others are maintained. In the following Sect. 5.2.3, we present four most important cases of such assumptions which can be combined together and are often considered in practice. While in general, most of the interest is devoted to the estimation of the $\underline{b}'$s, the $\underline{\beta}'$s or $\underline{\delta}$'s can receive an interesting interpretation in terms of errors of specification, allowing to test if the model corresponds to one given econometric specification or to another one. Actually, depending on the modeling of $\Phi(\underline{\beta})$ or $\Psi(\underline{\delta})$, there is often a trade-off between the accuracy of the estimation of $\underline{b}$ on the one hand, and its consistency and robustness to errors of specification in the other hand. Imposing more restrictions on the matrices $\Phi(\underline{\beta})$ or $\Psi(\underline{\delta})$, that is parameterizing them with a smaller vector $\underline{\beta}$ or $\underline{\delta}$, yields some efficiency gain in the estimation of $\underline{b}$. Conversely, the estimation of $\underline{b}$ to given errors of specification is more robust when imposing less restrictions on these matrices. In the simulation exercise presented in Sect. 5.5, we show that this can be indeed a crucial trade-off in small or midlle size panel data samples.

5.2.3 Four Important Cases

Let us consider the form of the estimating equations and of the restrictions on the Π matrix in four important usual cases: that of Correlated Effects (CE), Errors in Variables (EV), Weak Simultaneity (WS) and Lagged Dependent Variables (LDV). We consider them separately, but they can be combined together easily. These examples are also treated in more detail in Appendix B: Vector representation of the Chamberlain estimating equations.

5.2.3.1 Correlated Effects

In the case of correlated effects we assume that the past, present and future values of the explanatory variables x are not correlated with the (time varying) idiosyncratic disturbance u_{it}, but that they can be correlated with the individual effects μ_i. This implies that the Φ and the Ψ matrices have the same pattern with equal within-

[8] Note that when the estimating equations take the form of (5.11), identification of $\underline{b}$ and $\underline{\delta}$ depends on the particular form of $M(\underline{b})$ and $\Psi(\underline{\delta})$. When they take the form of (5.10) or that of (5.9), identification of $\underline{b}$ and $\underline{\beta}$ depends on the form of $M(\underline{b})$ and $\Phi(\underline{\beta})$, but also on $E(x_i x_i')$, and requires some specific conditions on $E(x_i x_i')$. See in the next subsection the cases of Errors in Variables (EV) and Weak Simultaneity (WS).

column coefficients, and thus can be parameterized by KT parameters β or δ.[9] We can simply write

$$\begin{aligned}\Phi(\underline{\beta}) &= E\left(\left(\mu_i \underline{l}\right)\underline{x}_i'\right) = \underline{l}E\left(\mu_i \underline{x}_i'\right) = \underline{l}\underline{\beta}' \\ \Psi(\underline{\delta}) &= E\left(\left(\mu_i \underline{l}\right)\underline{x}_i'\right)E\left(\underline{x}_i\underline{x}_i'\right)^{-1} = \underline{l}E\left(\mu_i \underline{x}_i'\right)E\left(\underline{x}_i\underline{x}_i'\right)^{-1} = \underline{l}\underline{\delta}'\end{aligned} \tag{5.12}$$

where $\underline{l}$ is the $(T \times 1)$ vector of 1's, $\underline{\beta}$ is the $(KT \times 1)$ vector of the covariances of μ_i and the x's, and $\underline{\delta}$ is the $(KT \times 1)$ vector of the coefficients of the linear projection of μ_i on the x's.

5.2.3.2 Errors in Variables

In the case of errors in variables we assume that the true model is

$$\underline{y}_i = M(b)\underline{x}_i^* + \underline{v}_i, \quad i = 1, \ldots, N$$

but that instead of x^* we only observe

$$\underline{x}_i = \underline{x}_i^* + \underline{e}_i$$

where the true $\underline{x}_i^*$ and the errors of measurement $\underline{e}_i$ are uncorrelated. We also assume for simplicity that the x^*'s are strictly exogenous (i.e., uncorrelated with the overall disturbance v) and that the errors of measurement themselves e are also uncorrelated with the v's. Denoting the covariance matrix of the measurement errors by $V_e = E(\underline{e}_i\underline{e}_i')$, we can then write

$$\Phi = E\left[\left(\underline{v}_i - M(\underline{b})\underline{e}_i\right)\underline{x}_i'\right] = -M(\underline{b})V_e \tag{5.13}$$

and

$$\Psi = -M(\underline{b})V_e E\left(\underline{x}_i\underline{x}_i'\right)^{-1}.$$

To identify the model it is necessary to make some simplifying assumptions on V_e. It is usually considered that measurement errors are serially uncorrelated (or so called "white noise"). We show in Appendix B that in this case the matrix $\Phi = -M(\underline{b})V_e$ in (5.13) has the simple form of the juxtaposition of K diagonal matrices, with (KT) possibly different diagonal coefficients. It follows from equation (5.9) that the parameters $\underline{b}$ are identifiable as long as for each k there is at least one l and two periods (s,t), with $s \neq t$, such that $E\left(x_{it}^{(k)}x_{is}^{(l)}\right) \neq 0$. Note, however, that the KT diagonal elements of the matrix $\Phi = -M(\underline{b})V_e$ are of the form $\sum_{l=1}^{K} b_l Cov\left(e_{it}^{(k)}, e_{it}^{(l)}\right)$, and thus it is only these KT functions of the $KT(K+1)/2$ parameters $\mathrm{Cov}\left(e_{it}^{(k)}, e_{it}^{(l)}\right)$ which are identifiable, not the individual

[9] Note that this corresponds simply to the fact that the linear projection of the disturbances v_{it} on the x's is the linear projection of the individual effects μ_i on the x's, and hence does not depend on t. Likewise $\mathrm{Cov}(v_{it}x_{is}) = \mathrm{Cov}(\mu_i x_{is})$ depends on s, not on t.

$\text{Cov}\left(e_{it}^{(k)}, e_{it}^{(l)}\right)$. This is why it is usually assumed that the measurement errors are not only serially uncorrelated, but also uncorrelated with each other, implying $\sum_{l=1}^{K} b_l \text{Cov}\left(e_{it}^{(k)}, e_{it}^{(l)}\right) = b_k E\left(e_{it}^{(k)2}\right)$, and thus allowing the estimation of the measurement errors variances $E\left(e_{it}^{(k)2}\right)$, and not only the estimation of the parameters of interest b's.

By contrast to Φ and $E(\underline{y}_i \underline{x}_i')$, Ψ and Π have a complicated pattern, involving leads and lags of the different x's, and they cannot be directly parameterized, irrespective of $E\left(\underline{x}_i \underline{x}_i'\right)^{-1}$:

$$\Pi = M(\underline{b}) + \Psi = M(\underline{b})[I - V_e E\left(\underline{x}_i \underline{x}_i'\right)^{-1}]. \tag{5.14}$$

It is easy to see that Correlated Effects and Errors in Variables can be considered jointly. Note that Correlated Effects alone take care of possible measurement errors that are constant over time.

5.2.3.3 Weak Simultaneity

"Weak Simultaneity" (WS), as we prefer to call it here, corresponds to the case of predetermination, or weak exogeneity, of the x variables, or of some of them. It allows for lagged effects of the y variable and possibly for contemporaneous two-ways effects by not assuming that the past and present idiosyncratic disturbances or shocks u_{is} can affect the current x's, but it assumes that future shocks do not. Note that we can equivalently say that past x 's are uncorrelated with the current shocks u_{is}. In this case, the identifying restrictions are

$$E\left(u_{is} x_{it}\right) = 0 \quad \text{for } s > t \tag{5.15}$$

and the matrix Φ has the characteristic pattern of a repeated upper triangular matrix which can be parameterized by $\left(KT\left(T+1\right)/2\right)$ parameters of secondary interest β. As previously, the parameterization of Ψ, and hence of Π follows from that of Φ and involves $E\left(\underline{x}_i \underline{x}_i'\right)^{-1}$.

Equations (5.15) are enough for identification when the explanatory variables x are correlated over time. To see this, we can assume for simplicity that $K = 1$, and write equations (5.9) for the couples of (s,t) indexes with $s > t$. We have $E\left(y_{is} x_{it}\right) = bE\left(x_{is} x_{it}\right) + E\left(u_{is} x_{it}\right) = bE\left(x_{is} x_{it}\right)$, showing that the parameter b will be identified as long as there is at least one (s,t), with $s > t$, such that $E\left(x_{is} x_{it}\right) \neq 0$.

Weak Simultaneity is usually combined with Correlated Effects. As considered here, it assumes either that the unobserved individual effects μ_i are uncorrelated with the x's or that they are equal to zero (i.e., $\mu_i = 0$).

In the case of Correlated Effects and Weak Simultaneity (CEWS), the identifying restrictions (5.15) become

$$\left[E\left(v_{is} x_{it}\right) - E\left(v_{is-1} x_{it}\right)\right] = E\left(\left(u_{is} - u_{is-1}\right) x_{it}\right) = 0 \quad \text{for } s > t+1.$$

Assuming for simplicity as above that $K = 1$, we see that the parameter b will be identified as long as there is at least one (s,t) with $s > t+1$ such that we have $E\left((x_{is} - x_{is-1})\ x_{it}\right) \neq 0$. This will fail, however, when the x's follow a random walk, or will not work well if they are very persistent (strongly autocorrelated).

Note finally that the case of Weak Simultaneity includes the possibility of errors in variables, if we assume that they are serially uncorrelated.

5.2.3.4 Lagged Dependent Variables

Let us also explicitly consider the case of an autoregressive model, in which Weak Simultaneity and Correlated Effects naturally arise from the presence of lagged dependent variables among the explanatory variables. Assuming a first order autoregressive model to keep computation simple, we can write it as:

$$y_{it} = \alpha y_{it-1} + \sum_k x_{it}^{(k)} \underline{b}_k + \mu_i + u_{it}$$

or in a vector format as:

$$\underline{y}_i = \alpha \underline{y}_{i(-1)} + M\left(\underline{b}\right) \underline{x}_i + \mu_i \underline{l} + \alpha y_{i0} \underline{l}_1 + \underline{u}_i$$

where $\underline{y}'_{i(-1)} = (0, y_{i1}, \ldots, y_{iT-1})$, $\underline{l}$ is the $(T \times 1)$ vector (1,1,...,1) and $\underline{l}'_1 = (1,0,\ldots,0)$. Using the $(T \times T)$ matrix L such that $\underline{y}_{i(-1)} = L\underline{y}_i$ (i.e., such that all the coefficients of the first subdiagonal are 1 and all the others are zeros), we can also write

$$[I - \alpha L]\underline{y}_i = M\left(\underline{b}\right) \underline{x}_i + \mu_i \underline{l} + \alpha y_{i0} \underline{l}_1 + \underline{u}_i. \tag{5.16}$$

Assuming that the x's can also be correlated with the fixed effect μ_i as well as with the initial (unknown) y_{i0}, but not with the shocks $\underline{u}_{it}$, and denoting respectively by $\sum_{s=1}^{T} \underline{\zeta}_s x_{is}$ the projection of y_{i0} on $\underline{x}_i$, and $\sum_{s=1}^{T} \underline{\delta}_s x_{is}$ the projection of μ_i on $\underline{x}_i$, we directly obtain from (5.16) the following set of estimating equations in term of the Π matrix and of the parameters of primary and secondary interest $\alpha, \underline{b}, \underline{\delta}$ and $\underline{\lambda} = \alpha \underline{\zeta}$

$$[I - \alpha L]\Pi = M\left(\underline{b}\right) + \underline{l}\underline{\delta}' + \underline{l}_1 \underline{\lambda}'. \tag{5.17}$$

Although they involve the Π matrix alone, these equations do not take the form of direct restrictions on Π (as in the Correlated Effects case), but of a more general implicit function $f\left(\Pi, \underline{\theta}\right) = 0$, where $\underline{\theta}' = (\alpha, \underline{b}, \underline{\delta}, \underline{\lambda})$. They can also be transformed in terms of direct restrictions by left-multiplying them by $[I - \alpha L]^{-1}$, but this leads to more complex nonlinear relations between Π and the different parameters of interest.

5.2.4 Restrictions on the Covariance Matrix of the Disturbances

The restrictions on the Π matrix considered so far arise from assumptions on the covariances between the overall disturbances $\underline{v}_i$ and the $\underline{x}_i$'s, by imposing a given structure to the matrices Ψ or Φ and allowing them to be more or less simply parameterized. We have not made any assumptions on the structure of the covariance matrix $\Sigma = E(\underline{v}_i \underline{v}_i')$ of these disturbances. The Chamberlain method estimators of the parameters of primary interest $\underline{b}$ and of secondary interest $\underline{\delta}$ or $\underline{\beta}$ (and their standard errors) are thus robust to non constant year variances and to any kind of serial correlation in these disturbances.[10]

However, we may be interested in making some simplifying assumptions on the form of Σ. For example we may want to test that $\underline{v}_i$ has (indeed) a pure error component structure $\mu_i + u_{it}$, or that its time varying component u_{it} is not homoscedastic, or that it is not serially uncorrelated, but generated by an autoregressive (AR) or moving average (MA) process. Such assumptions also give rise to restrictions on the covariances matrices of $\underline{y}_i$ and $\underline{x}_i$, and hence can be used to improve the efficiency of the estimators.

More precisely, consider the case when Σ can be expressed in terms of additional parameters $\underline{\lambda}$ (of dimension less than $T(T+1)/2$, the number of individual year variances and covariances in Σ). We can write the following relations:

$$\begin{aligned} \Sigma(\underline{\lambda}) &= E[\underline{v}_i \underline{v}_i'] = E[(\underline{y}_i - M(\underline{b})\underline{x}_i)(\underline{y}_i - M(\underline{b})\underline{x}_i)'] \\ &= E[\underline{y}_i \underline{y}_i'] - M(\underline{b})E[\underline{x}_i \underline{y}_i'] - E[\underline{y}_i \underline{x}_i']M(\underline{b})' + M(\underline{b})E[\underline{x}_i \underline{x}_i']M(\underline{b})'. \end{aligned} \tag{5.18}$$

Denoting the residual of the projection of $\underline{y}_i$ on $\underline{x}_i$ by $\underline{w}_i = \underline{y}_i - \Pi \underline{x}_i$ and its covariance matrix by $V_w = E(\underline{w}_i \underline{w}_i')$, we can also write

$$\Sigma(\underline{\lambda}) = V_w + \Phi(\underline{\beta})E[\underline{x}_i \underline{x}_i']^{-1}\Phi(\underline{\beta})' \tag{5.19}$$

or

$$\Sigma(\underline{\lambda}) = V_w + \Psi(\underline{\delta})E[\underline{x}_i \underline{x}_i']\Psi(\underline{\delta})' \tag{5.20}$$

The previous relations (5.18) are of course equivalent to the relations (5.19) or (5.20), but the later relations are probably a better way to write them. If we can assume that the y's and x 's are normally distributed (or approximately so), we know that the estimates of Π, $E(\underline{x}_i \underline{x}_i')$ and V_w are independent (or approximately so).

In principle these two sets of relations have a similar status to that of the previous covariance conditions (5.10) and (5.11). Both sets impose restrictions between the parameters of interest $\underline{b}$, $\underline{\beta}$ or $\underline{\delta}$, and $\underline{\lambda}$, and the coefficients of Π and covariances

[10] The way in which the Φ or Ψ matrices are parameterized may imply, however, some a priori restrictions on the covariance matrix of the disturbances. Note that if the overall disturbances include additional disturbance terms uncorrelated with the x's, their covariance matrix is modified, while the parameterization of Φ and of Ψ remains unchanged. Random (uncorrelated) coefficients, for example, can be the source of such additional disturbances.

of $\underline{y}_i$ and $\underline{x}_i$ (to be estimated in the first stage of the Chamberlain method). Both can be used as estimating equations for the parameters of interest in the second stage ALS.[11]

In practice, however, important differences must be noted. First, one may not be willing to impose any structure on the covariance matrix of the disturbances Σ, because one is not confident or simply not interested in doing so.[12] Second the restrictions on Σ are more complicated. They necessarily involve the moments $E(\underline{x}_i\underline{x}_i')$ and $E(\underline{y}_i\underline{y}_i')$ (or V_w) in addition to the Π matrix, and they are nonlinear in the parameters of interest. In Appendix C: Manipulation of equations and parameters in the ALS framework, we show that the nonlinearity problem can be partially overcome if we proceed in three stages instead of two (that is if we impose the restrictions on Σ only in a third stage).

5.2.5 A Generalization of the Chamberlain Method

In this section we provide an extended view of the Chamberlain methodology which applies to more general models than the basic regression setting just considered. In Appendix A: An extended view of the Chamberlain method, we show how it applies more specifically to simultaneous equations models, vector autoregressions (VAR), and endogenous attrition models.

Assuming it is linear in its disturbances, an econometric model can always be written as a set of stochastic equations expressing these disturbances $\underline{v}_i$ in terms of the variables $\underline{z}_i' = (\underline{y}_i', \underline{x}_i')$,without making an explicit distinction between the dependent and explanatory variables $\underline{y}_i$ and $\underline{x}_i$. We can thus write:

$$\begin{array}{l} \underline{v}_i = \underline{v}_i(\underline{z}_i, \underline{\theta}) = A(\underline{\theta})\underline{z}_i + \underline{d} \qquad \text{with} \\ E(\underline{v}_i) = 0 \text{ and } E(\underline{v}_i\underline{v}_i') = \Omega(\underline{\theta}), \end{array}$$

leading to the following moment conditions or estimating equations:

$$\begin{array}{l} A(\underline{\theta})E(\underline{z}_i) + \underline{d} = 0 \\ \Omega(\underline{\theta}) = A(\underline{\theta})E(\underline{z}_i\underline{z}_i')A(\underline{\theta})' + \underline{d}\underline{d}' \end{array} \tag{5.21}$$

---

[11] The parameterization of $\Sigma(\underline{\lambda})$ of Σ will usually be linear. This is the case for example when the disturbances have an error component structure, and when the time varying disturbances follow an MA process. However in some interesting cases the parameterization may be non linear: for example, when the time varying disturbances follow an AR(1) process $u_{it} = \rho u_{it-1} + w_{it}$, where w_{it} is a white noise. However, in this situation the model can be transformed by the quasi-difference operator $[I - \rho L]$ into an autoregressive regression, leading to the estimating equations $[I - \rho L]\Pi = [I - \rho L]M(\underline{b}) + \Phi(\underline{\beta})E(\underline{x}_i\underline{x}_i')^{-1}$ where $\Phi\left(\underline{\beta}\right)$ is now the parameterization of the covariance matrix $E[\underline{w}_i\underline{x}_i']$ and the variance matrix $\Sigma = E(\underline{w}_i\underline{w}_i')$ can be linearly parameterized.

[12] On the contrary, one may be particularly interested in the serial correlation structure of the dependent variable *per se* without being willing to make any assumption on the Π matrix. In this case, we have $M(\underline{b}) = \Pi$ with $\Phi(\underline{\beta}) = \Psi(\underline{\delta}) = 0$, so that $V_w = \Sigma(\lambda)$ are the only equations to be considered.

where A is a matrix parameterized by a vector of parameters $\underline{\theta}$ of (primary and secondary) interest, $\underline{d}$ is a vector of constants, usually period or year constants, and Ω is the covariance matrix of the disturbances also parameterized by $\underline{\theta}$.

Writing a model in such a general form usually involves many parameters with very little substantive content, in which we are not really interested. For example, in all the cases considered so far, the matrix $E(\underline{x}_i\underline{x}_i')$ of second order moments of the explanatory variables is unconstrained and implicitly parameterized by its $KT(KT+1)/2$ individual elements. The problem is thus how to eliminate the set, or only a subset, of parameters $\underline{\theta}_s$ of secondary interest, in such a way that it does not affect the efficiency of the ALS estimator for the remaining subset of parameters $\underline{\theta}_p$. The intuitive solution is to solve for $\underline{\theta}_s$ in an appropriate subset of equations, as function of the $\underline{\theta}_p$ and the moments of the variables; then to substitute them in the remaining equations, so that they only include $\underline{\theta}_p$ and the moments; and finally to proceed to the estimation on these equations.

It can be shown that the ALS estimators of the parameters of interest on the transformed and reduced set of equations can be as efficient as those on the full set of equations, as long as the numbers of eliminated parameters and eliminated equations are equal (Crépon, Kramarz and Trognon, 1998). This result can be very useful in practice. For example, one can get rid of all the constant terms $\underline{d}$ in the equations (as long as there are no constraints on them) by discarding the first order conditions and centering the variables at their overall means, or at their period or year means. The reduced set of moment conditions simply becomes:

$$\Omega(\underline{\theta}) = A(\underline{\theta})E(\underline{z}_i\underline{z}_i')A(\underline{\theta})' - A(\underline{\theta})E(\underline{z}_i)E(\underline{z}_i')A(\underline{\theta})' \tag{5.22}$$

$$= A(\underline{\theta})V_{\underline{z}_i}A(\underline{\theta})' \tag{5.23}$$

where $V_{\underline{z}_i} = E\left(\underline{z}_i\underline{z}_i'\right) - E\left(\underline{z}_i\right)E\left(\underline{z}_i'\right)$.

An even more obvious case is when the parameters to be eliminated are separately entering one equation only. Then one has simply to drop such equations. We can thus discard all the equations corresponding to the matrix $E(\underline{x}_i\underline{x}_i')$ of second order moments of the explanatory variables, if the model does not imply any restrictions involving it (contrary, for example, to the VAR model we consider in Appendix A). Likewise, we can eliminate the equations corresponding to the matrix $E(\underline{y}_i\underline{y}_i')$ of the second order moments of the dependent variable if no restrictions are made on the covariance matrix $E(\underline{v}_i\underline{v}_i')$ of the disturbances (contrary to what we do in the previous Sect. 5.2.4 and again in the case of the VAR model in Appendix A).

5.2.6 The Vector Representation of the Chamberlain Estimating Equations

In practice, in order to apply Asymptotic Least squares (ALS) to the Chamberlain estimating equations, we have to write these equations in vector form rather than in matrix form. A systematic method to do this is to apply a "Vec" operator, which

simply stacks one above the others the columns of a matrix. This operator has some convenient properties that makes such transformations easy and powerful. If $A = (\underline{c}_1 \cdots \underline{c}_K)$, by definition

$$\mathrm{Vec}(A) = \begin{pmatrix} \underline{c}_1 \\ \vdots \\ \underline{c}_K \end{pmatrix}, \tag{5.24}$$

and it can be shown in particular that if the matrix A of dimension $L \times K$ is the external product $(\underline{v}_1 \underline{v}_2')$ of the two column vectors $\underline{v}_1$ and $\underline{v}_2$ of dimension L and K respectively, then the column vector $\mathrm{Vec}(A)$ of dimension LK is equal to the Kronecker product of $\underline{v}_1$ by $\underline{v}_2$: $\mathrm{Vec}(A) = \mathrm{Vec}(\underline{v}_1 \underline{v}_2') = \underline{v}_2 \otimes \underline{v}_1$.

In Appendix B we recall some other properties of this operator and apply them to show that the vector representation of the four important specifications of the linear regression model (5.5) considered in Sect. 5.2.3, can take the general form:

$$\underline{\pi} - H(\underline{m})\underline{\gamma} = 0, \tag{5.25}$$

where $\underline{\pi} = \mathrm{Vec}(\Pi')$ is the column vector of the matrix Π stacked by rows, $\underline{\gamma}$ is the parameter column vector of parameters of primary and secondary interest, and H is a matrix function of $\underline{m}$, where $\underline{m}$ and the dimension of H vary with the specification. While the matrix H is constant in the case of Correlated Effects (CE), it depends on $\underline{m} = E(\underline{x}_i \underline{x}_i')$ in the cases of Errors in Variables (EV) and of Weak Simultaneity (WS). In the case of a Lagged Dependent Variable (LDV) specification, H is a function of $\underline{m} = \underline{\pi}$ itself, implying an implicit relation between $\underline{\pi}$ and $\underline{\gamma}$.

5.2.7 The Estimation of Matrix Π

The auxiliary parameters entering the Chamberlain estimating equations are the moments of the variables, or are functions of them as the coefficients of Π; they are also functions of the covariance matrix V_w of the residuals $\underline{w}_i$ of the linear projection of $\underline{y}_i$ on $\underline{x}_i$, if restrictions are imposed on the serial correlation of the disturbances. We not only have to estimate these moments, but also the asymptotic variances of their limiting distribution, since these variances play an important role in the practical implementation of ALS estimators in the second stage of the Chamberlain method. As we shall see in the next Sect. 5.3, they are necessary to compute consistent estimators for the standard errors of the estimated parameters of interest, and they are also needed in order to compute the optimal (efficient) ALS estimators and to perform specification tests.

The estimation of the moments and their asymptotic variances is straightforward. Using the notation $\underline{z}_i' = (\underline{y}_i', \underline{x}_i')$, we want to estimate the vector of moments $\underline{m} = E(\underline{m}_i)$ and its covariance matrix $V_m = Var(\underline{m}_i)$, where $\underline{m}_i = \mathrm{Vec}(\underline{z}_i \underline{z}_i') = (\underline{z}_i \otimes \underline{z}_i)$, or more precisely $\underline{m}_i = D(\underline{z}_i \otimes \underline{z}_i)$, D being a selection matrix which keeps only the

different cross-products from $\text{Vec}(\underline{z}_i\underline{z}_i')$.[13] The expectation $\underline{m}$ and covariance V_m in the population are estimated by the corresponding empirical mean $\widehat{\underline{m}}$ and covariance $\widehat{V}_{\widehat{\underline{m}}}$ in the sample:

$$\widehat{\underline{m}} = \frac{1}{N}\sum_{i=1}^{N}\underline{m}_i \qquad \text{and} \qquad \widehat{V}_{\widehat{\underline{m}}} = \frac{1}{N}\sum_{i=1}^{N}(\underline{m}_i - \widehat{\underline{m}})(\underline{m}_i - \widehat{\underline{m}})'.$$

Direct application of the weak law of large numbers and of the central limit theorem tells that under very general conditions $\widehat{\underline{m}}$ is a consistent estimator of $\underline{m} = E(\underline{m}_i)$ with a normal asymptotic distribution of covariance $V_{\widehat{\underline{m}}}$, which is itself consistently estimated by $\widehat{V_{\widehat{\underline{m}}}}$:

$$\widehat{\underline{m}} \xrightarrow{P} E(\underline{m}_i) \quad \text{with} \quad \sqrt{N}(\widehat{\underline{m}} - E(\underline{m}_i)) \xrightarrow{D} N(0, V_{\widehat{\underline{m}}}) \quad \text{and} \quad \widehat{V_{\widehat{\underline{m}}}} \xrightarrow{P} V_{\widehat{\underline{m}}}.$$

5.2.7.1 Estimation of Matrix Π Alone

The vector $\underline{\pi}$ ($=$ vec (Π')), formed by stacking the column vectors of the transposed matrix Π, can be directly estimated as the vector of coefficients in the overall system of the T stacked year regressions of $\underline{y}_i$ on $I_T \otimes \underline{x}_i'$:

$$\underline{y}_i = (I_T \otimes \underline{x}_i')\underline{\pi} + \underline{w}_i.$$

The Generalized Least Squares (GLS) estimator $\widehat{\underline{\pi}}$ of $\underline{\pi}$ is given by

$$\begin{aligned}\widehat{\underline{\pi}} &= \overline{(I_T \otimes \underline{x}_i')'(I_T \otimes \underline{x}_i')}^{-1}\overline{(I_T \otimes \underline{x}_i')\underline{y}_i} \\ &= I_T \otimes \overline{\underline{x}_i\underline{x}_i'}^{-1}\overline{\underline{y}_i \otimes \underline{x}_i}\end{aligned} \tag{5.26}$$

where a bar over an expression $h(z_i)$ stands for the empirical mean over the sample (i.e., $\overline{h(z_i)} = 1/N\sum_i h(z_i)$).[14] This estimator follows asymptotically a normal distribution and its asymptotic covariance matrix $V_{\widehat{\pi}}$ is equal to:

$$V_{\widehat{\pi}} = [I_T \otimes E(\underline{x}_i\underline{x}_i')^{-1}]E(\underline{w}_i\underline{w}_i' \otimes \underline{x}_i\underline{x}_i')[I_T \otimes E(\underline{x}_i\underline{x}_i')^{-1}]. \tag{5.27}$$

---

[13] Using the operator Vech for a symetric matrix, one can also write $m_i = \text{Vech}(\underline{z}_i\underline{z}_i')$. See Appendix B.

[14] Note that the GLS estimator $\widehat{\underline{\pi}}$ is identical to the estimator obtained by stacking as a column vector the T separate OLS estimators of the row vectors of coefficients $(\pi_{t.})$ in the T separate year regressions $y_{it} = \pi_{t.}\underline{x}_i + w_{it}$ (or (5.2)), since these T separate regressions have the same regressors $\underline{x}_i$. The GLS asymptotic covariance matrix $V_{\widehat{\pi}}$ in the case of homoscedastic errors also coincides with the corresponding asymptotic covariance matrix estimated on the basis of the T separate OLS estimators. However, it is also consistently estimated in the case of heteroscedastic errors (see below), while the latter is not.

$V_{\widehat{\pi}}$ can be consistently estimated by $\widehat{V}_{\widehat{\pi}}$ obtained by replacing in (5.27) the expectations $E(\underline{x}_i\underline{x}_i')^{-1}$ by the sample averages $\overline{\underline{x}_i\underline{x}_i'}^{-1}$ and the errors $\underline{w}_i$ by the estimated residuals $\widehat{\underline{w}}_i = \underline{y}_{.i} - (I_T \otimes \underline{x}_i')\widehat{\pi}$.

$\widehat{V}_{\widehat{\pi}}$ is robust to the heteroscedasticity of the errors $\underline{w}_i$ (White 1980). Note, however, that the middle term $E(\underline{w}_i\underline{w}_i' \otimes \underline{x}_i\underline{x}_i')$ in the expression of $V_{\widehat{\pi}}$ includes moments of the fourth order, which can be poorly estimated. Under the assumption of homoscedasticity, this term simplifies to:

$$E(\underline{w}_i\underline{w}_i' \otimes \underline{x}_i\underline{x}_i') = E(E(\underline{w}_i\underline{w}_i'|\underline{x}_i) \otimes \underline{x}_i\underline{x}_i') = E(\underline{w}_i\underline{w}_i') \otimes E(\underline{x}_i\underline{x}_i')\,,$$

and $V_{\widehat{\pi}}$ also simplifies to:

$$V_{\pi}^{c} = E(\underline{w}_i\underline{w}_i') \otimes E(\underline{x}_i\underline{x}_i')^{-1}\,, \tag{5.28}$$

which now involves only moments of the second order that can be more precisely estimated.

5.2.7.2 Joint Estimation of Matrix Π and Other Relevant Moments

The Chamberlain estimating equations, as we have seen in Sects. 5.2.2–5.2.4, often include, in addition to the coefficients of matrix Π, other relevant moments such as $E(\underline{x}_i\underline{x}_i')$ in the cases of the Error in Variables and Weak Simultaneity specifications, and $E(\underline{w}_i\underline{w}_i')$ when simplifying assumptions are imposed on the structure of covariance matrix of the disturbances. In such cases, the column vector of all auxiliary parameters to be estimated in the first stage of the Chamberlain method is not only $\underline{\pi}$ but $\underline{\pi}^e = (\underline{\pi}', \underline{m}_w', \underline{m}_x')'$ where $\underline{m}_w = \text{Vec}(E(\underline{w}_i\underline{w}_i') = E(\underline{w}_i \otimes \underline{w}_i)$ and $\underline{m}_x = \text{Vec}(E(\underline{x}_i\underline{x}_i') = E(\underline{x}_i \otimes \underline{x}_i)$.

While $\underline{m}_x$ can be directly estimated by the corresponding sample average $\widehat{\underline{m}}_x$, this is not so for $\underline{m}_w$ since the residuals $\underline{w}_i$ are not observed and have first to be estimated themselves as $\widehat{\underline{w}}_i = \underline{y}_{.i} - (I_T \otimes \underline{x}_i')\widehat{\underline{\pi}}$. However, the estimator computed by simply taking the sample average of the Kronecker product of the estimated residuals: $\widehat{\underline{m}}_w = \overline{\widehat{\underline{w}}_i \otimes \widehat{\underline{w}}_i}$ has the same asymptotic limiting behavior as if these residuals were exactly known. It can thus be shown that $\widehat{\underline{\pi}^e} = (\widehat{\underline{\pi}}', \widehat{\underline{m}}_w', \widehat{\underline{m}}_x)'$ has the following asymptotic joint normal distribution:

$$\sqrt{N}\begin{pmatrix} \widehat{\underline{\pi}} - \underline{\pi} \\ \overline{\widehat{\underline{w}}_i \otimes \widehat{\underline{w}}_i} - \underline{m}_w \\ \overline{\underline{x}_i \otimes \underline{x}_i} - \underline{m}_x \end{pmatrix} \xrightarrow{D} N\left(0, \begin{pmatrix} V_{11} & & \\ V_{21} & V_{22} & \\ V_{31} & V_{23} & V_{33} \end{pmatrix}\right)$$

with

$$\begin{aligned} V_{11} &= V_{\widehat{\pi}} \text{ as in (5.27)}, \\ V_{21} &= E\left[\underline{w}_i\underline{w}_i' \otimes (\underline{w}_i\underline{x}_i' E(\underline{x}_i\underline{x}_i')^{-1})\right], \end{aligned}$$

$$
\begin{aligned}
V_{31} &= E\left[\underline{x}_i\underline{w}_i' \otimes (\underline{x}_i\underline{x}_i' E(\underline{x}_i\underline{x}_i')^{-1})\right], \\
V_{22} &= E\left[\underline{w}_i\underline{w}_i' \otimes (\underline{w}_i\underline{w}_i')\right] - \underline{m}_w\underline{m}_w', \\
V_{32} &= E\left[\underline{x}_i\underline{w}_i' \otimes (\underline{x}_i\underline{w}_i')\right] - \underline{m}_x\underline{m}_w', \\
V_{33} &= E\left[\underline{x}_i\underline{x}_i' \otimes (\underline{x}_i\underline{x}_i')\right] - \underline{m}_x\underline{m}_x'.
\end{aligned}
$$

As indicated before for $V_{11} = V_{\widehat{\pi}}$, all the asymptotic covariance matrices V_{sl} involve moments of the fourth order and thus can be poorly estimated. However, also as before, their expression can be substantially simplified under the assumption that the distribution of the residuals $\underline{w}_i$ conditional on $\underline{x}_i$ is homoscedastic and normal. Under this assumption, the covariance matrix V_{11} of $\widehat{\underline{\pi}}$ is the one given in (5.28), and V_{22} is only function of the moments of second order in $\underline{m}_w$, while V_{21}, V_{31} and V_{23} are zero matrices. Likewise, if the normality assumption can be extended to the explanatory variables x, V_{33} can be expressed as a similar function of the moments of second order in $\underline{m}_x$.

5.3 Asymptotic Least Squares

5.3.1 ALS Estimation

The Chamberlain second stage estimators are based on the so called Minimum Distance method (Malinvaud, 1970, Chamberlain, 1982) or also known as the Asymptotic Least Squares method (Gourieroux, Monfort and Trognon, 1985). This method applies to situations in which the n_θ parameters of (primary and secondary) interest $\underline{\theta}$ to be estimated are related by n_g estimating equations to n_μ auxiliary parameters $\underline{\mu}$, which have already been consistently estimated. From now on in Sects. 5.3 and 5.4, we shall usually speak of *the* parameter $\underline{\theta}$ and *the* parameter $\underline{\mu}$ and we shall note, whenever needed, their true values by $\underline{\theta}^0$ and $\underline{\mu}^0$, the assumption being that these true values verify exactly the n_g estimating equations $g(\underline{\theta}^0, \underline{\mu}^0) = 0$.

5.3.1.1 Basic Result

Let $\widehat{\underline{\mu}}$ be a consistent and asymptotically normal estimator of the auxiliary parameter $\underline{\mu}$, computed in a first stage on a sample of size N, and let $V_{\widehat{\mu}}$ be its asymptotic covariance matrix $V_{\widehat{\mu}}$, that is $\sqrt{N}(\widehat{\underline{\mu}} - \underline{\mu}^0) \xrightarrow{D} N(0, V_{\widehat{\mu}})$. The principle of the ALS method is to choose a $\widehat{\underline{\theta}}$ such that estimating equations $g(\widehat{\underline{\theta}}, \widehat{\underline{\mu}})$ are as close as possible to zero. Since the dimension n_g of g is usually larger than that n_θ of $\underline{\theta}$, it is impossible to have exactly $g(\widehat{\underline{\theta}}, \widehat{\underline{\mu}}) = 0$ and hence $\widehat{\underline{\theta}}$ is computed by minimizing a weighted quadratic sum of the g's.

More precisely, defining $\widehat{\underline{\theta}}(S)$ as:

$$
\widehat{\underline{\theta}}(S) = \text{ Arg min}_\theta \left\{ g(\underline{\theta}, \widehat{\underline{\mu}})' S_N g(\underline{\theta}, \widehat{\underline{\mu}}) \right\}, \tag{5.29}
$$

where $S = (S_N)_{N=1,\ldots,\infty}$ is a sequence of weight matrices possibly depending on the sample, and assuming that the weight matrix S_N converges in probability at the rate N to a given matrix S_0 and that the functions g verify some identifiability and regularity conditions,[15] it can be shown that $\widehat{\underline{\theta}}(S)$ is a consistent and asymptotically normal estimator of the true parameter $\underline{\theta}^0$, that is:

$$\sqrt{N}(\widehat{\underline{\theta}}(S) - \underline{\theta}^0) \xrightarrow{D} N(0, V_{\widehat{\underline{\theta}}(S)})$$

with

$$V_{\widehat{\underline{\theta}}(S)} = \left[\frac{\partial g'}{\partial \underline{\theta}} S_0 \frac{\partial g}{\partial \underline{\theta}'}\right]^{-1} \frac{\partial g'}{\partial \underline{\theta}} S_0 \frac{\partial g}{\partial \underline{\mu}'} V_{\widehat{\underline{\mu}}} \frac{\partial g'}{\partial \underline{\mu}} S_0 \frac{\partial g}{\partial \underline{\theta}'} \left[\frac{\partial g'}{\partial \underline{\theta}} S_0 \frac{\partial g}{\partial \underline{\theta}'}\right]^{-1}. \tag{5.30}$$

In this formula, the partial derivative matrices $\frac{\partial g}{\partial \underline{\theta}'}$ and $\frac{\partial g}{\partial \underline{\mu}'}$ are evaluated at $\underline{\theta}^0, \underline{\mu}^0$. A consistent estimator $\widehat{V}_{\widehat{\underline{\theta}}(S)}$ can be obtained by computing them at $\widehat{\underline{\theta}}, \widehat{\underline{\mu}}$ and by replacing $V_{\widehat{\underline{\mu}}}$ by a consistent estimator $\widehat{V}_{\widehat{\underline{\mu}}}$. Obviously the choice of the weight matrix affects the asymptotic behavior of the ALS estimator. In Sect. 5.3.2, we show that there exists an optimal choice such that the corresponding ALS estimator is asymptotically efficient.

5.3.1.2 Application to the Chamberlain Approach

The implementation of ALS is simplified when the estimating equations are linear in the parameter of interest $\underline{\theta}$ and thus the objective function is a simple quadratic function of $\underline{\theta}$. This applies in the Chamberlain framework when the restrictions on the Π matrix implied by the modeling assumptions can be written in a vector form as:

$$\underline{\pi}^0 = H(\underline{m}^0)\underline{\theta}^0 \tag{5.31}$$

where $\underline{m}^0$ is a vector of second order moments of the variables. As shown in Sect. 5.2.6 and Appendix B, this applies in particular for the four main specifications of correlated effects, errors in variables, weak simultaneity and lagged dependent variable and their combinations.

When this is the case, we can derive the following explicit expression for the solution $\widehat{\underline{\theta}}(S)$ of (5.29):

$$\widehat{\underline{\theta}}(S) = [H(\widehat{\underline{m}})' S_N H(\widehat{\underline{m}})]^{-1} H(\widehat{\underline{m}})' S_N \widehat{\underline{\pi}} = P(S_N, \widehat{\underline{m}})\widehat{\underline{\pi}}. \tag{5.32}$$

[15] The regularity conditions are that g is twice continuously differentiable and that $\frac{\partial g'}{\partial \underline{\theta}} S_0 \frac{\partial g}{\partial \underline{\theta}'}$ is invertible when evaluated at the true $\underline{\theta}^0$. The identifiability condition is that $g(\underline{\theta}, \underline{\mu}^0) = 0$ implies $\underline{\theta} = \underline{\theta}^0$. This condition requires that $n_g \geq n_{\underline{\theta}}$.

The implementation of ALS is further simplified when the H matrix is constant, as in the case of the correlated effects specification. In this case, the asymptotic covariance matrix of $\widehat{\underline{\theta}}(S)$ given by (5.30) simply becomes:

$$V_{\widehat{\underline{\theta}}(S)} = [H'S_0H]^{-1}H'S_0V_{\underline{\widehat{\pi}}}S_0H[H'S_0H]^{-1} = P(S_0)V_{\underline{\widehat{\pi}}}P(S_0)'. \tag{5.33}$$

To obtain both $\widehat{\underline{\theta}}$ and a consistent estimator $V_{\widehat{\underline{\theta}}(S)}$ of its covariance matrix, it is thus enough to know $\widehat{\underline{\pi}}$ and a consistent estimator of its covariance matrix $V_{\widehat{\underline{\pi}}}$.[16]

If the H matrix is not a constant, as in the case of three other specifications, the formula defining $V_{\widehat{\underline{\theta}}(S)}$ is more complicated:

$$V_{\widehat{\underline{\theta}}(S)} = P(S_0, \underline{m}^0)V(\widehat{\pi}, \widehat{m})P(S_0, \underline{m}^0)' \tag{5.34}$$

where instead of simply being $V_{\widehat{\underline{\pi}}}$, the matrix $V(\widehat{\pi}, \widehat{m})$ is

$$V(\widehat{\pi}, \widehat{m}) = \left\{ \left[I - \frac{\partial H\underline{m}^0}{\partial \underline{m}}\underline{\theta}^0 \right] V_{\widehat{\pi}, \widehat{m}} \left[I - \frac{\partial H\underline{m}^0}{\partial \underline{m}}\underline{\theta}^0 \right] \right\}. \tag{5.35}$$

In the case of errors in variables or weak simultaneity where $\underline{m}$ is $E(\underline{x}_i\underline{x}_i')$, it is thus necessary to compute the empirical second order moments in $\overline{\underline{x}_i\underline{x}_i'}$ in addition to $\widehat{\underline{\pi}}$ to obtain $\widehat{\underline{\theta}}(S)$. But it is also necessary to compute the covariance matrices of these estimators, which involves moments of the fourth order, in order to estimate the asymptotic covariance $V_{\widehat{\underline{\theta}}(S)}$. Neglecting this complication and using formula (5.33) instead of (5.34) will result in a biased estimator for the asymptotic covariance $V_{\widehat{\underline{\theta}}(S)}$.

5.3.2 The Optimal ALS Estimator

The asymptotic properties of the different ALS estimators $\widehat{\underline{\theta}}(S)$ depend on their limiting weight matrix S_0. If $W = \frac{\partial g}{\partial \underline{\mu}'}V_{\underline{\mu}}\frac{\partial g'}{\partial \underline{\mu}}$ is invertible, there is an optimal choice $S_0^* = W^{-1}$ leading to an asymptotically efficient estimator (meaning that for any weight matrix S_0 different from S_0^*, there exists a symmetric positive matrix Δ such that: $V_{\widehat{\underline{\theta}}(S_0)} = V_{\widehat{\underline{\theta}}(S_0^*)} + \Delta$).[17] The asymptotic covariance matrix for the optimal ALS estimator $\widehat{\underline{\theta}}(S^*)$ thus simplifies as follows:

---

[16] Note that the formulas of (5.32) and (5.33) giving $\widehat{\underline{\theta}}$ and $V_{\widehat{\underline{\theta}}(S)}$ are the formulas of the weighted least squares estimator of $\underline{\theta}_0$ (with weight matrix S_N) in the linear regression model $\widehat{\underline{\pi}} = H(\widehat{\underline{m}})\underline{\theta}_0 + \varepsilon$ providing a first order approximation to the estimating equations (5.31), with $\varepsilon = -\frac{\partial H(\underline{m}^0)}{\partial \underline{m}}(\underline{\theta}_0)(\widehat{\underline{m}} - \underline{m}_0) + (\widehat{\underline{\pi}} - \underline{\pi}_0)$.

[17] The condition that $W = \frac{\partial g}{\partial \underline{\mu}'}V_{\underline{\mu}}\frac{\partial g'}{\partial \underline{\mu}}$ requires that there is no solution to the equation $(\partial g'/\partial \underline{\mu})\underline{v} = 0$, which, in turn, requires that the dimension of $\underline{\mu}$ exceeds or equal that of g : $n_\mu \geq n_g$.

$$V_{\widehat{\theta}(S_0^*)} = \left[\frac{\partial g'}{\partial \underline{\theta}} S_0^* \frac{\partial g}{\partial \underline{\theta}'}\right]^{-1} = \left[\frac{\partial g'}{\partial \underline{\theta}}\left[\frac{\partial g}{\partial \underline{\mu}'} V_{\underline{\mu}} \frac{\partial g'}{\partial \underline{\mu}}\right]^{-1} \frac{\partial g}{\partial \underline{\theta}'}\right]^{-1}. \tag{5.36}$$

When the Chamberlain estimating equations take the form of (5.31): $\underline{\pi}^0 = H(\underline{m}^0)\underline{\theta}^0$ and if the matrix H is constant as in the case of the correlated effects specification, the asymptotic covariance matrix of the optimal ALS estimator simplifies further as:

$$V_{\widehat{\theta}(S_0^*)} = \left[H' V_{\widehat{\underline{\pi}}}^{-1} H\right]^{-1}. \tag{5.37}$$

Note that in the case the optimal ALS estimator is the Generalized Least Squares estimator of $\underline{\theta}$ in the linear regression equation $\widehat{\underline{\pi}} = H\underline{\theta} + \varepsilon$ (see footnote 16 in the preceding Sect. 5.3.1.2).

5.3.2.1 Implementation of the Optimal ALS Estimator

The practical implementation of the optimal ALS estimator $\widehat{\underline{\theta}}(S^*)$ is actually very similar to that of the Generalized Least Square estimator. Since the optimal weight matrix is generally unknown being a function of the true parameter $\underline{\theta}^0$ (and of the variance of the estimated auxiliary parameter $\widehat{\underline{\mu}}$), it has to be generally performed in two steps. In the first step a consistent ALS estimate $\widehat{\underline{\theta}}$ is computed using an arbitrary weight matrix (and the consistent estimates of $\widehat{\underline{\mu}}$ and $\widehat{V}_{\widehat{\underline{\mu}}}$ already obtained). In the second step, this estimate and the previous estimates of $\widehat{\underline{\mu}}$ and $\widehat{V}_{\widehat{\underline{\mu}}}$ are used to derive a consistent estimator $\widehat{S}_0^* = \left[\frac{\partial g}{\partial \underline{\mu}'}(\widehat{\underline{\theta}}, \widehat{\underline{\mu}})\widehat{V}_{\mu}\frac{\partial g'}{\partial \underline{\mu}}(\widehat{\underline{\theta}}, \widehat{\underline{\mu}})\right]^{-1}$ of the optimal weight matrix S_0^* and compute the estimator $\widehat{\underline{\theta}}(\widehat{S}_0^*)$. Since $\widehat{S}_0^*$ converges in probability to S_0^*, the estimator $\widehat{\underline{\theta}}(\widehat{S}_0^*)$ obtained in this second step is asymptotically efficient.

It is not always necessary to implement a two step procedure in the context of the Chamberlain framework. When the estimating equations take the form $\underline{\pi}^0 = f(\underline{\theta}^0)$ the covariance matrix of $\widehat{\underline{\pi}}$ is already the W matrix and there is no need for a first step, and the asymptotic covariance matrix of the optimal $\widehat{\underline{\theta}}(S_0^*)$ is given directly by the following expression which generalizes (5.37):

$$V_{\widehat{\underline{\theta}}(S_0^*)} = \left[\frac{\partial f'}{\partial \underline{\theta}} V_{\widehat{\underline{\pi}}}^{-1} \frac{\partial f}{\partial \underline{\theta}'}\right]^{-1}. \tag{5.38}$$

One advantage of the ALS is its flexibility in allowing nested estimation and testing. Let us assume that $\underline{\theta}^0$(of dimension n_θ) is itself subject to restrictions and can be expressed more parsimoniously in terms of a parameter $\underline{\gamma}^0$ (of smaller dimension $n_\gamma < n_\theta$) as: $\underline{\theta}^0 - q(\underline{\gamma}^0) = 0$. Two estimation procedures are possible: a direct one in which the estimation of $\underline{\gamma}$ is performed on the basis of the set of estimating equations

$g(q(\underline{\gamma}^0), \underline{\mu}^0)$, and an indirect one in which an estimated $\widehat{\underline{\theta}}$ is obtained as previously and then used as an auxiliary parameter to estimate $\underline{\gamma}$ on the basis of the reduced set of estimating equations: $\underline{\theta}^0 - q(\underline{\gamma}^0) = 0$. It can be shown that the direct and indirect estimators of $\underline{\gamma}$ are asymptotically equivalent if the optimal ALS is implemented in both cases.

5.3.2.2 Finite Sample Properties of the Optimal ALS Estimator

The optimal properties of the two step ALS estimator are asymptotic ones with the sample size N. In practice they may require a very large sample size N to hold precisely enough. Simulation experiments (performed in the related GMM context by Arellano and Bond, 1991, on small samples of size $N = 100$) tend to show that the one step estimators may be practically as good as the two steps optimal estimator. Moreover, these simulation experiments indicate that the estimated standard errors of the one step estimators are satisfactory, while the estimated standard errors of the two steps estimators can be downward biased. These results have been confirmed in the context of nonlinear models by Bertscheck and Lechner (1995).

Such poor performance is related to the estimation of the optimal weight matrix. This matrix is the inverse of $W = \frac{\partial g}{\partial \underline{\mu}'} V_{\underline{\mu}} \frac{\partial g'}{\partial \underline{\mu}}$. Note that W (not its inverse) enters in (5.30) from which the standard errors of the first step estimates are obtained. Thus, the problem is twofold: to have a good estimator of W and to have a good estimator of its inverse. The reason why W may be poorly estimated is that it involves moments of the fourth order which, for a sample of a given size, are less precisely estimated than moments of second order. W inverse can also be poorly estimated even if W is not, since a small error in W can lead to a magnified error in its inverse. This happens when W is "badly conditioned", that is when the "condition number" of this matrix is high, where the condition number of matrix A is defined as:

$$c(A) = ||A||.||A||^{-1} = \frac{\max \text{ (eigenvalues of } A)}{\min \text{ (eigenvalues of } A)}.^{18}$$

A limiting case is when W is not invertible (and the condition number is infinite). This can happen when the number of observations is not large enough relatively to the size of W. In the case where $W = \widehat{V}_{\widehat{\underline{\pi}}} = [I_T \otimes E(\underline{x}_i \underline{x}_i')^{-1}][E(\underline{w}_i \underline{w}_i' \otimes \underline{x}_i \underline{x}_i')][I_T \otimes E(\underline{x}_i \underline{x}_i')^{-1}]$, W is invertible only if $(\underline{w}_i \otimes \underline{x}_i)(\underline{w}_i \otimes \underline{x}_i)'$ is also invertible, which requires N to be larger than KT^2 (i.e., the dimension of the column vector $\underline{w}_i \otimes \underline{x}$). For example, if $T = 10$ and $K = 2$ the number of observations N must be greater than 200.

[18] The condition number is always greater than one. Large values indicate that in the computation of the inverse of a matrix $A + E$ the relative error $\left\|(A+E)^{-1} - A^{-1}\right\| / \left\|A^{-1}\right\|$ can be high compared with the relative error $\|E\| / \|A\|$ (see Stewart, 1973). The condition number is an upper bound for this relative error.

5.3.3 Specification Testing in the ALS Framework

Generally, the number of estimating equations n_g is much larger than the dimension n_θ of the parameter of interest $\underline{\theta}$. A specification test of the model based on the fact that the equations must be verified at the true $\underline{\theta}^0$ (i.e., a test for overidentifying restrictions) can be implemented. The intuition behind the test is to check whether an appropriately weighted quadratic form of the residuals in the estimating equations is small, implying that these residuals are all small indeed. More precisely, under the null hypothesis that the estimating equations are compatible (i.e., $\exists\ \underline{\theta}^0 \,/\, g(\underline{\theta}^0, \underline{\mu}^0) = 0$), it can be shown that the weighted quadratic form of the residuals $\zeta(S)$ converges in distribution towards a χ^2 with $(n_g - n_{\underline{\theta}})$ degrees of freedom

$$\zeta(S) = N g(\widehat{\underline{\theta}}(S), \widehat{\underline{\mu}}) \left[V(g(\widehat{\underline{\theta}}(S), \widehat{\underline{\mu}})) \right]^{-} g(\widehat{\underline{\theta}}(S), \widehat{\underline{\mu}}) \xrightarrow{P} \chi^2(n_g - n_{\underline{\theta}}), \tag{5.39}$$

where $[V(g(\widehat{\theta}(S), \widehat{\mu}))]^{-}$ is a generalized inverse of the asymptotic covariance matrix of the residuals of the estimating equations at the estimated values of the parameters, $\widehat{\underline{\theta}}$ and $\widehat{\underline{\mu}}$.[19] Note that $[V(g(\widehat{\theta}(S), \widehat{\mu}))]^{-}$ is not the given weight matrix S used in the estimator $\widehat{\underline{\theta}}$, except when $S = S^*$ is the optimal weight matrix (as explained below in Sect. 5.3.3.1). Thus, when implementing the optimal ALS in a two step procedure, the objective function for the first step is not a valid test statistic; it is only valid in the optimal ALS. In order to perform the test after the first step one has to recompute the objective function using $[V(g(\widehat{\theta}(S), \widehat{\mu}))]^{-}$ instead of S.

It is also important to note that the test statistics $\zeta(S)$ are asymptotically equivalent under the null hypothesis for all weight matrices S.[20] Therefore, the asymptotic properties of the tests does not depend on whether an arbitrary ALS estimator or the optimal one has been used.

5.3.3.1 Andrews' Problem

The actual implementation of the specification test in (5.38) raises a difficult problem, known as Andrews' problem (Andrews, 1985). The covariance matrix V of the residuals has the following form

$$\begin{aligned} V(S,W) &= V(g(\widehat{\underline{\theta}}(S), \widehat{\underline{\mu}})) \\ &= \left[I - \frac{\partial g}{\partial \underline{\theta}'} \left[\frac{\partial g'}{\partial \underline{\theta}} S \frac{\partial g}{\partial \underline{\theta}'} \right]^{-1} \frac{\partial g'}{\partial \underline{\theta}} S \right] W \left[I - \frac{\partial g}{\partial \underline{\theta}'} \left[\frac{\partial g'}{\partial \underline{\theta}} S \frac{\partial g}{\partial \underline{\theta}'} \right]^{-1} \frac{\partial g'}{\partial \underline{\theta}} S \right]' \end{aligned}$$

[19] V is the asymptotic covariance matrix of the residuals $g(\widehat{\theta}(S), \widehat{\mu})$ multiplied by $\sqrt{N}$, and therefore the proper covariance matrix of the residuals is V/N. This is why the test statistic $\zeta(S)$ is written with a factor N.

[20] This means that the difference between any two test statistics $(\zeta(S_1) - \zeta(S_2))$ converges towards zero in probability. See Gourieroux and Monfort (1989), and also Newey (1985) for a related issue on specification tests in GMM.

with $W = \frac{\partial g}{\partial \underline{\mu}'} V_{\underline{\mu}} \frac{\partial g'}{\partial \underline{\mu}}$. By construction, this is the covariance matrix of the projection of the vector $g(\underline{\theta}_0, \widehat{\underline{\mu}})$ (whith asymptotic covariance matrix W) on the space orthogonal to the subspace generated by the columns of the matrix $\frac{\partial g}{\partial \underline{\theta}'}$, i.e.,

$$\left[I - \frac{\partial g}{\partial \underline{\theta}'} \left[\frac{\partial g'}{\partial \underline{\theta}} S \frac{\partial g}{\partial \underline{\theta}'} \right]^{-1} \frac{\partial g'}{\partial \underline{\theta}} S \right] g(\underline{\theta}^0, \widehat{\underline{\mu}})$$

in the metric defined by the weight matrix S. Clearly this is not an invertible matrix. Thus it is necessary to compute the weight matrix used in the test as a generalized inverse of V. The Andrews' problem arises from the fact that the V matrix is not known and has to be estimated, and from the non continuity of the generalized inverse operation. It results that a generalized inverse of a consistent estimator of V is not necessarily a consistent estimator of a generalized inverse V^-.

One way to solve this problem is to find a specific generalized inverse of $\widehat{V}(g(\widehat{\underline{\theta}}(S), \widehat{\underline{\mu}}))$ which is a continuous function of $\widehat{\theta}$, $\widehat{\mu}$ and of $\widehat{V}_{\widehat{\underline{\mu}}}$. There are two cases in which this can be done. The first is when the test is based on the optimal ALS estimator. The second corresponds to a re-formulation of the test based on a reduced form of the estimating equations.

1. It can be shown that the optimal weight matrix W^{-1} used to implement the optimal ALS estimator is a particular generalized inverse of the covariance matrix V.[21] Since $\widehat{W} = \frac{\partial g}{\partial \underline{\mu}'}(\widehat{\underline{\mu}}, \widehat{\underline{\theta}}) \widehat{V_{\underline{\mu}}} \frac{\partial g'}{\partial \underline{\mu}}(\widehat{\underline{\mu}}, \widehat{\underline{\theta}})$ is a consistent estimator of W and since the inverse is a continuous operator, $\widehat{W}^{-1}$ is a consistent estimator of W^{-1}. Therefore the test can be implemented, by using the objective function of the optimal ALS as a test statistic:

$$N g(\widehat{\underline{\theta}}^*, \widehat{\underline{\mu}})' S_N^* g(\widehat{\underline{\theta}}^*, \widehat{\underline{\mu}}) \xrightarrow{D} \chi^2(n_g - n_{\underline{\theta}}) \tag{5.40}$$

where $\widehat{\underline{\theta}}^*$ is the optimal ALS estimator, and $S_N^* = \widehat{W}^{-1}$ the optimal weight matrix.

2. Assume that the parameter of interest $\underline{\theta}$ can be solved out in terms of the auxiliary parameter $\underline{\mu}$ using $n_{\underline{\theta}}$ of the n_g estimating equations. After the elimination of $\underline{\theta}$ the remaining $n_g - n_{\underline{\theta}}$ equations $h(\underline{\mu}) = 0$ must be verified at the true value $\underline{\mu}_0$. These equations $h(\underline{\mu}^0) = 0$, are the direct expression of the overidentifying restrictions of the model and they can be simply tested with the statistic

$$N h(\widehat{\underline{\mu}})' \left(\frac{\partial h}{\partial \underline{\mu}} \widehat{V_{\underline{\mu}}} \frac{\partial h'}{\partial \underline{\mu}} \right)^{-1} h(\widehat{\underline{\mu}}) \xrightarrow{D} \chi^2(n_g - n_{\underline{\theta}}). \tag{5.41}$$

[21] We have to verify that: $VW^{-1}V = V$. If we write $V = [I-P]W[I-P]'$, with $P = \frac{\partial g}{\partial \underline{\theta}'} \left[\frac{\partial g'}{\partial \underline{\theta}} W^{-1} \frac{\partial g}{\partial \underline{\theta}'} \right]^{-1} \frac{\partial g'}{\partial \underline{\theta}} W^{-1}$, it is straightforward to see that $[I-P]W = W[I-P]'$. The result then follows from the fact that $[I-P]$ is a projector (idempotent), which implies $[I-P]^2 = [I-P]$.

It can be shown that this statistic is asymptotically equivalent to the previous test statistics.[22]

5.4 The Equivalence of the GMM and the Chamberlain Methods

In this section we show how the extended view of the Chamberlain method given in Sect 5.2.5 can be reinterpreted in terms of the Generalized Method of Moments (GMM). More precisely we show that the relations between the moments and the parameters of interest used as the Chamberlain ALS estimating equations can also be taken as orthogonality conditions which can be used to implement GMM estimators.

Starting with the general linear model

$$\underline{v}_i = v_i(\underline{z}_i, \underline{\theta}_0) = A(\underline{\theta}_0)\underline{z}_i,$$

and with

$$E(\underline{v}_i) = 0, \text{ and } E(\underline{v}_i \underline{v}_i') = \Omega(\underline{\theta}_0).$$

we can write in vector form the following moment conditions:

$$\begin{array}{l} A(\underline{\theta}_0)E(\underline{z}_i) = 0 \\ D[A(\underline{\theta}_0) \otimes A(\underline{\theta}_0]E(\underline{z}_i \otimes \underline{z}_i) = D \text{ Vec } (\Omega(\underline{\theta}_0)) \end{array} \tag{5.42}$$

where the selection matrix D keeps only the elements of the symetric matrix on the diagonal or below. Since these expressions are linear in the moments, they can be equivalently written as orthogonality conditions

$$E(h(\underline{z}_i, \underline{\theta}_0)) = 0, \tag{5.43}$$

where

$$h(\underline{z}_i, \underline{\theta}) = \begin{pmatrix} A(\underline{\theta})z_i \\ D[A(\underline{\theta}) \otimes A(\underline{\theta})]\underline{z}_i \otimes \underline{z}_i - D \text{ Vec } (\Omega(\underline{\theta})) \\ \cdot \end{pmatrix} \tag{5.44}$$

5.4.1 A Reminder on the GMM

Before proceeding, we recall briefly the principle of GMM estimation. As just indicated, GMM is based on the orthogonality conditions expressing that a given

[22] In fact, it is one of them corresponding to the special choice of a weight matrix that only weights the equations used to eliminate $\underline{\theta}$. The problem of the generalized inverse is solved here as in the case of the optimal estimator, because it is possible to find a specific generalized inverse which is a continuous function of the parameters and variances and therefore can be consistently estimated by replacing these parameters and variances by their estimates.

function h of the parameter $\underline{\theta}$ and the variables z has a zero expectation at the true value $\underline{\theta}_0$ of the parameter. The principle of estimation is to minimize a quadratic form of the empirical counterpart of these orthogonality conditions with respect to $\underline{\theta}$

$$\widetilde{\underline{\theta}} = \widetilde{\underline{\theta}}(S) = \underset{\underline{\theta}}{\text{Arg min}} \left[\frac{1}{N}\sum_i h(\underline{z}_i, \underline{\theta})\right]' S_N \left[\frac{1}{N}\sum_i h(\underline{z}_i, \underline{\theta})\right],$$

where $S = [S_N]_{N=1,\ldots,\infty}$ is a sequence of weight matrices, and $\widetilde{\underline{\theta}} = \widetilde{\underline{\theta}}(S)$ is the resulting GMM estimator of $\underline{\theta}$.

Under some identifiability and regularity conditions on h, it can be shown that whatever the choice of the sequence of weight matrices S, provided it converges in probability to a weight matrix S_0, the GMM estimator $\widetilde{\underline{\theta}}$ converges in probability to the true $\underline{\theta}^0$ and is asymptotically normally distributed, with an asymptotic covariance matrix $V_{\widetilde{\underline{\theta}}(S)}$ of $\widetilde{\underline{\theta}}$ depending on S_0. More precisely, we have

$$\sqrt{N}(\widetilde{\underline{\theta}} - \underline{\theta}_0) \to N(0, V_{\widetilde{\underline{\theta}}(S)}),$$

with

$$V_{\widetilde{\underline{\theta}}(S)} = \left[G(\underline{\theta}_0)' S_0 G(\underline{\theta}_0)\right]^{-1} G(\underline{\theta}_0)' S_0 W S_0 G(\underline{\theta}_0) \left[G(\underline{\theta}_0)' S_0 G(\underline{\theta}_0)\right]^{-1},$$

where $G(\underline{\theta}) = E\left(\frac{\partial h}{\partial \underline{\theta}'}(\underline{z}_i, \underline{\theta})\right)$ and $W = V(h(\underline{z}_i, \underline{\theta}_0)) = E(h(\underline{z}_i, \underline{\theta}_0) h(\underline{z}_i, \underline{\theta}_0)')$. This matrix $V_{\widetilde{\underline{\theta}}(S)}$ is a function of both $G(\underline{\theta}_0)$ and W, which are unknown, but can be estimated consistently by:

$$\widetilde{G}(\widetilde{\underline{\theta}}) = \frac{1}{N}\sum_i \frac{\partial h(z_i, \widetilde{\underline{\theta}})}{\partial \underline{\theta}'} \text{ and } \widetilde{W} = \frac{1}{N}\sum_i h(\underline{z}_i, \widetilde{\underline{\theta}}) h(\underline{z}_i, \widetilde{\underline{\theta}})'.$$

As for the ALS (see Sect. 5.3.2), there is a special choice S_0^* of the limit of the weight matrices S that makes the corresponding estimator optimal (based on the same set of orthogonality conditions). This corresponds to $S_0^* = W^{-1}$, the inverse of the covariance matrix of the orthogonality conditions. In this case the asymptotic covariance matrix of $\widetilde{\underline{\theta}}$ becomes

$$V_{\widetilde{\underline{\theta}}(S_0^*)} = \left[G(\underline{\theta}_0)' W^{-1} G(\underline{\theta}_0)\right]^{-1}.$$

Since the W matrix is unknown, the optimal GMM estimator cannot be directly implemented. As for the ALS it is necessary to proceed in two steps and to compute in a first step a consistent estimator $\widetilde{W}$ based on a GMM estimator with an arbitrary weight matrix. Note that the considerations concerning the small sample properties of the optimal ALS estimators also apply to the optimal GMM estimators (see Sect. 5.3.2.2).

5.4.2 Equivalence of the GMM and the Chamberlain Methods

Denoting $\underline{Z}_i$ the vector formed by the independent elements of $\underline{z}_i$ and $\left(\underline{z}_i \otimes \underline{z}_i\right)$, the ALS estimating equations (5.41) and the GMM orthogonality conditions (5.42) can be rewritten as:

$$B(\underline{\theta}_0)E(\underline{Z}_i) - C(\underline{\theta}_0) = 0$$

and

$$E(B(\underline{\theta}_0)\underline{Z}_i - C(\underline{\theta}_0)) = 0.$$

It is easy to see that if we use the same weight matrix S, both estimators are not only asymptotically equivalent but also numerically identical. Indeed the ALS estimator $\widehat{\underline{\theta}} = \widehat{\underline{\theta}}(S)$ results from the minimization of

$$\left(\left[B(\underline{\theta})\frac{1}{N}\sum_i \underline{Z}_i - C(\underline{\theta})\right]' S \left[B(\underline{\theta})\frac{1}{N}\sum_i \underline{Z}_i - C(\underline{\theta})\right]\right)$$

and the GMM estimator $\widetilde{\underline{\theta}} = \widetilde{\underline{\theta}}(S)$ from that of

$$\left(\frac{1}{N}\sum_i [B(\underline{\theta})\underline{Z}_i - C(\underline{\theta})]' S \frac{1}{N}\sum_i [B(\underline{\theta})\underline{Z}_i - C(\underline{\theta})]\right).$$

The two estimators are identical, since the two objective functions are obviously the same:

$$\frac{1}{N}\sum_i [B(\underline{\theta})\underline{Z}_i - C(\underline{\theta})] = B(\underline{\theta})\frac{1}{N}\sum_i \underline{Z}_i - C(\underline{\theta}).$$

It follows that the optimal estimators are also identical. Indeed, we can verify that the optimal ALS weight matrix $S^*_{\text{ALS}} = [B(\underline{\theta}_0)V(\underline{Z}_i)B(\underline{\theta}_0)']^{-1}$ is obviously equal to the optimal GMM weight matrix $S^*_{\text{GMM}} = [V(B(\underline{\theta}_0)\underline{Z}_i - C(\underline{\theta}_0))]^{-1}$.

In practice, however, the optimal weight matrices S^*_{ALS} and S^*_{GMM} have to be estimated, and this can be done in several ways. Numerical differences between the optimal ALS and GMM estimators can thus arise in small samples. Let us mention three reasons why this is actually happening. The first one is just a simple matter of computation, while the other two are related to the different ways in which the Chamberlain method and the GMM are implemented.

Assume that we dispose of a first step estimate $\widehat{\underline{\theta}}_1 = \widetilde{\underline{\theta}}_1$ obtained with either one of the two methods for a given weight matrix. The ALS optimal weight matrix is computed as $\widehat{S}^*_{\text{ALS}} = [B(\widehat{\underline{\theta}}_1)\widehat{V}(\underline{Z}_i)B(\widehat{\underline{\theta}}_1)']^{-1}$. For GMM, let us denote by $\widetilde{\underline{R}}_i$ the residual of the orthogonality condition for the observation i: $\widetilde{\underline{R}}_i = B(\widetilde{\underline{\theta}}_1)\underline{Z}_i - C(\widetilde{\underline{\theta}}_1)$. Since $E(\widetilde{\underline{R}}_i) = 0$, there are two consistent estimators of the optimal GMM weight matrix:

$$\widetilde{S}^{*1}_{\text{GMM}} = \overline{\widetilde{\underline{R}}_i \widetilde{\underline{R}}_i'}$$

$$\widetilde{S}^{*2}_{\text{GMM}} = \overline{\widetilde{\underline{R}}_i \widetilde{\underline{R}}_i'} - \overline{\widetilde{\underline{R}}_i}\,\overline{\widetilde{\underline{R}}_i'} = \overline{\left(\widetilde{\underline{R}}_i - \overline{\widetilde{\underline{R}}_i}\right)\left(\widetilde{\underline{R}}_i - \overline{\widetilde{\underline{R}}_i}\right)'}$$

where a bar over an expression stands for its empirical mean over the sample. It is usually $\widetilde{S}^{*1}_{\mathrm{GMM}}$ which is computed, while it is $\widetilde{S}^{*2}_{\mathrm{GMM}}$ which equals $\widehat{S}^{*}_{\mathrm{ALS}}$.

The results on the elimination of parameters of secondary interest presented in Appendix C can be extended to the GMM.[23] If the number of discarded orthogonality conditions is the same as the number of eliminated parameters, there is no loss of asymptotic efficiency in the estimation of the remaining parameters; if it is larger, there is a loss of asymptotic efficiency, but consistency is preserved. Contrary to the Chamberlain approach, the usual practice of GMM amounts to considering orthogonality conditions which only involve the parameters of primary interest (and implicitly eliminating the parameters of secondary interest). If all such orthogonality conditions are taken into account, both the optimal Chamberlain and GMM estimators are equivalent but not identical, since they are not computed on the basis of the same weight matrix S. If only a subset of them is used (for example, the ones corresponding to the most obvious valid instruments), the GMM estimator is less efficient than the Chamberlain estimator.

The GMM always requires a two step estimation to implement the optimal estimator. In the Chamberlain method this is not always the case. When the estimating equations take the form of a direct linear restriction on $\underline{\pi}$, the optimal weight matrix is simply the inverse of the covariance matrix of the $\underline{\pi}$ estimator (see 5.37). It is also important to note that in the Chamberlain approach it is possible to keep the same weight matrix when considering a sequence of nested specifications, while in the GMM case a weight matrix must be computed for each different specification. This may be in practice an advantage of the Chamberlain approach.

5.4.3 Equivalence in Specific Cases

We have just shown the general equivalence of the Chamberlain and GMM estimators when all restrictions on all (first and second order) moments of the variables are considered. It is straightforward to see that the equivalence holds as well when we only focus on the conditions involving the joint moments of the dependent and explanatory variables, and any given subset of moment conditions. The ALS estimators based on the estimating equations as written in Sect. 5.2.2 is

$$\underline{\Pi}_0 = M(\underline{b}_0) + \Phi(\underline{\beta}_0) E\left(\underline{x}_i \underline{x}_i'\right)^{-1}$$

or

$$E(\underline{y}_i \underline{x}_i') = M(b_0) E\left(\underline{x}_i \underline{x}_i'\right) + \Phi(\underline{\beta}_0)$$

which are clearly equivalent to the GMM estimators based on the orthogonality conditions

$$E\left[\underline{y}_i \underline{x}_i' - M(b_0)\underline{x}_i \underline{x}_i' - \Phi(\underline{\beta}_0)\right] = 0.$$

[23] See Crépon, Kramarz and Trognon (1998).

It may be of some interest to be more specific and illustrate the general equivalence of the GMM and Chamberlain methods in the three important cases of Correlated Effects, Errors in Variables and Weak Simultaneity.

5.4.3.1 Correlated Effects

Let us suppose for convenience that we only have one explanatory variable. Thus we have T^2 estimating equations, or orthogonality conditions, for one parameter of primary interest. We also have T parameters of secondary interest corresponding to the correlated effects. From the expression of Φ given in Sect. 5.2.2 the T^2 orthogonality conditions can be written as

$$E\left(\underline{y}_i\underline{x}_i' - b\underline{x}_i\underline{x}_i' - \underline{l}\underline{\beta}'\right) = 0,$$

where the β's are the covariances between the individual effects and the yearly x's. It is easy to see that since the $T \times T$ matrix $\Phi = \underline{l}\underline{\beta}'$ is constant in columns, premultiplying it by the $(T-1)\times T$ difference matrix Δ results in a $(T-1)\times T$ zero matrix.[24] Thus premultiplying the T^2 orthogonality conditions by Δ eliminates the T parameters of secondary interest $\underline{\beta}$, and gives $(T-1)\times T$ transformed orthogonality conditions:

$$E(\Delta\left(\underline{y}_i\underline{x}_i' - b\underline{x}_i\underline{x}_i'\right)) = E((\Delta\underline{y}_i - b\Delta\underline{x}_i)\underline{x}_i') = E((\Delta\underline{u}_i)\underline{x}_i') = 0.$$

This new set of orthogonality conditions simply expresses that the different year x's (in levels) can be used as instruments for the model after transforming it in first differences. They are clearly equivalent to the original set of conditions, since we have lost exactly T of them in eliminating the T parameters of secondary interest $\underline{\beta}$.

5.4.3.2 Errors in Variables

Assuming like in Sect. 5.2.3 that we have serially uncorrelated errors in variables e_{it} (but with possible varying variances $E\left(e_{it}^2\right)$) and combining them with correlated effects, the orthogonality conditions can be written as

$$E(\underline{y}_i\underline{x}_i' - b\underline{x}_i\underline{x}_i') = \begin{bmatrix} \beta_1+\lambda_1 & \beta_2 & \dots & \beta_T \\ \beta_1 & \beta_2+\lambda_2 & & \beta_T \\ \vdots & & \ddots & \\ \beta_1 & \beta_2 & & \beta_T+\lambda_T \end{bmatrix}$$

[24] See footnote 48 in Sect. 5.8.2 for the definition of the difference matrix Δ.

where $\beta_t = \text{cov}\ (\mu_i x_{it})$ as in the CE case, and $\lambda_t = -bE\left(e_{it}^2\right)$. We have now $2T$ parameters of secondary interest, and we are thus looking for $\left(T^2 - 2T\right)$ transformed orthogonality conditions only involving the parameter of primary interest $\underline{b}$. If we transform the model in first differences, and consider the year levels of x as potential instruments, clearly the past values lagged by two years and more and the future values are valid instruments, while the present values and the past values lagged by only one year are not since

$$\begin{aligned} E\left(\Delta v_{it} x_{is}\right) &= E\left(\left(\left(u_{it} - be_{it}\right) - \left(u_{it-1} - be_{it-1}\right)\right)\left(x_{is}^* + e_{is}\right)\right) \\ &= 0 \text{ if } s \neq t \text{ and } s \neq (t-1) \\ &= -bE\left(e_{is}^2\right) \text{ if } s = t \\ &= bE\left(e_{is}^2\right) \text{ if } s = (t-1) \end{aligned}$$

We are thus obtaining $(T-1)(T-2) = T^2 - 3T + 2$ orthogonality conditions that involve only the parameter $\underline{b}$, but we are still missing $(T-2)$ of them. These can be obtained by taking the second differences of the model, and instrumenting them by the in–between year levels of x. Clearly we have $E\left(\Delta v_{it} x_{it}\right) = -E\left(\Delta v_{it+1} x_{it}\right)$, and thus

$$E\left(\left(v_{it+1} - v_{it-1}\right) x_{it}\right) = 0.$$

These new $(T-2)$ equations are by construction independent of the preceding ones.

The total set of $T(T-2)$ orthogonality conditions are those considered by Griliches and Hausman (1986); they provide estimates as efficient as the ones of the Chamberlain method which is based on the T^2 estimating equations and gives estimates of both the parameter of primary interest and the $2T$ parameters of secondary interest.

5.4.3.3 Weak Simultaneity

In this case it is assumed that the current shocks are uncorrelated with the past values of the explanatory variable x (although they may affect its present and future values). It is easy to see that the matrix Φ must be upper triangular (see Sect. 5.2.3.3). Combining correlated effects and weak simultaneity, we have the T^2 orthogonality conditions

$$E\left(\underline{y}_i \underline{x}_i' - b\underline{x}_i \underline{x}_i'\right) = \begin{bmatrix} \beta_1 + \lambda_{11} & \beta_2 + \lambda_{12} & \dots & \beta_T + \lambda_{1T} \\ \beta_1 & \beta_2 + \lambda_{22} & & \beta_T + \lambda_{2T} \\ \vdots & & \ddots & \\ \beta_1 & \beta_2 & & \beta_T + \lambda_{TT} \end{bmatrix}$$

where there is now $T\left(T+1\right)/2$ parameters of secondary interest $\lambda_{st} = \text{cov}\ \left(u_{is}, x_{it}\right)$ for $s \leq t$ in addition to the T previous ones $\beta_t = \text{cov}\ \left(\mu_i x_{it}\right)$ for correlated effects.

However, in total, there are only $(T(T+1)/2)+(T-1)$ parameters, since only T of the $(T+1)$ parameters $\beta_T, \lambda_{1T}, \ldots, \lambda_{TT}$ can be identified (from the T covariances of $\underline{y}$ and the last year level x_{iT}), and we need $T^2 - T(T+1)/2 - (T-1) = \left(T^2 - 3T + 2\right)/2 = (T-1)(T-2)/2$ orthogonality conditions in terms of the parameter of interest only. These are exactly provided by instrumenting the model transformed in first differences with past levels of x lagged by two years and more:

$$E(\Delta u_{it} x_{is}) = E(u_{it} x_{is}) - E(u_{it-1} x_{is}) = 0 \quad \text{if} \quad s < (t-1).^{25}$$

5.4.3.4 Restriction on the Covariance Matrix of the Disturbances

Finally, it is worthwhile to consider also the case in which assumptions can be made about the covariance structure of the disturbances. Considering for example the case in which we assume an error components structure, we know that:

$$E(v_{is} v_{it}) = \sigma_\mu^2 + \sigma_{ut}^2 \delta_{t=s}, \tag{5.45}$$

where σ_{ut}^2 is the variance of the random shock u_{it} at time t. This corresponds to $(T(T+1)/2)$ new equations, to $(T+1)$ new parameters: $\sigma_\mu^2, \sigma_{u1}^2, \ldots, \sigma_{uT}^2$, and thus to $((T-2)(T+1)/2)$ supplementary orthogonality conditions in terms of the parameters of primary interest. We can derive them from the equations (5.45) by writing:

$$E(v_{it} v_{is}) = E(v_{it-1} v_{is}), \quad \text{if} \quad s < (t-1),$$

or:

$$E(\Delta u_{it} u_{is}) = 0, \quad \text{if} \quad s < (t-1),$$

and thus:

$$\begin{aligned} E(\Delta u_{it} y_{is}) &= E(\Delta u_{it}(b x_{is} + \mu_i + u_{is})) \\ &= bE(\Delta u_{it} x_{is}) + E(\Delta u_{it} \mu_i) + E(\Delta u_{it} u_{is}) = 0, \quad \text{if} \quad s < (t-1). \end{aligned}$$

These new $(T-1)(T-2)/2$ orthogonality conditions are simply expressing that the past values of y lagged by two years or more are valid instruments. The $(T-2)$ missing orthogonality conditions are less straightforward to derive (see Ahn and Schmidt 1995). They can be deducted by writing $E(\Delta u_{it} u_{it}) = \sigma_{ut}^2$ and $E(\Delta u_{it} u_{it-1}) = \sigma_{ut-1}^2$, implying the $(T-2)$ conditions:

$$E((u_{it+1} - u_{it-1}) u_{it}) = 0,$$

which can be rewritten as:

[25] Note that if we can assume stationarity, we can obtain an additional set of orthogonality conditions $E(u_{it} \Delta x_{is}) = 0$ for $s < t-1$. See last paragraph of Sect. 5.6.2, and Arellano and Bover (1995) and Blundell and Bond (1998).

$$bE\left(\left(u_{it+1}-u_{it-1}\right)x_{it}\right)+E\left(\left(u_{it+1}-u_{it-1}\right)y_{it}\right)=0.$$

In the case of Weak Simultaneity, these additional orthogonality conditions are nonlinear in the parameters of interest $\underline{b}$, but in the case of both Correlated Effects and Errors in Variables, they simply become:

$$E\left[\left(u_{it+1}-u_{it-1}\right)y_{it}\right]=0,$$

expressing that the in–between year levels of y can be used as valid instruments for the model transformed in second differences.

5.5 Monte Carlo Simulations

To give a feeling of how the Chamberlain method and the GMM perform in practice, we conducted some plausibly calibrated Monte-Carlo simulation experiments (see Tables 5.1 and 5.2). We consider the consistency and efficiency of the different estimators (see Tables 5.3–5.5), for simulated panel data samples of different lenght ($T=3$ and 6) and size ($N=100,400$ and 1600), in different "scenarios" corresponding to two main types of specification errors: Correlated Effects and Errors in Variables. We are also concerned with the consistency of the estimated standard errors (see Tables 5.6–5.8) and the performance of specification tests (see Tables 5.9 and 5.10). But first of all, let us provide some indications necessary to understand the design of the simulation experiments and thus the validity and limits of our results (for more details see Appendix E: Design of simulation experiments).

5.5.1 Design of the Simulation Experiments

Taking the simple linear regression model $y_{it}=\alpha x_{it}+(\mu_i+u_{it})$ with only one explanatory variable, we consider three basic "scenarios". The first scenario is that of Non Correlated Errors (NCE), in which the explanatory variable x is uncorrelated with both disturbance terms μ_i and u_{it}. The second one is that of Correlated Effects (CE) where the variable x is correlated with the individual effect μ_i, but remains uncorrelated with u_{it}. The first scenario thus corresponds to the standard basic case, while the second is usually regarded as more realistic since it takes advantage of the avaibility of panel data to control for potential unobserved correlated individual effects. Our third scenario combines the Correlated Effects and Errors in Variables cases (CE + EV) and can be considered as even more realistic. It assumes that the true explanatory variable x^* in the model is not observed and that the observed variable x is measured with a random measurement error, resulting in an additional error term, say φ_{it}, in the model and implying a correlation of the current x_{it} with the current φ_{it}, but not with the future and past φ_{is} ($s>t$ and $s<t$). We have also experimented with a fourth scenario combining Correlated Effects and Weak Simultaneity (CE + WS), as well as with scenarios with EV and WS only. These scenarios do

not provide much more insight than the three considered here, and we do not report their results.

We calibrated the experiments so that we can reproduce some of the features found in real data sets, in particular when estimating production functions on firm panel data, as in Mairesse (1990) and Griliches and Mairesse (1998), with y and x measuring respectively the log of the firm labor productivity and the log of the firm capital to labor ratio. Normalizing the (average) true value of the parameter of primary interest α to be 0.5, we assume that the variances of the (simulated) explanatory variable x and of the (simulated) disturbances $(\mu_i + u_{it})$ are of the same order of magnitude, normalizing them to be both 1. We also assume that most of the variability arises from the cross–sectional dimension of the data: that is, generating x as the sum of a between component ξ_i and a within component ξ_{it}, we take $\mathrm{Var}(\xi_i) = 0.8$ and $\mathrm{Var}(\xi_{it}) = 0.2$, and similarly we choose $\mathrm{Var}(\mu_i) = 0.8$ and $\mathrm{Var}(u_{it}) = 0.2$. Note that in addition we assume that the within component (ξ_{it}) of x is highly serially correlated according to a first order autocorrelation process with parameter 0.7, while we maintain the assumption that the usual errors u_{it} are not (auto)correlated. This implies that the past and future values of x can be used as valid and effective instruments in the case of the CE and CE+EV scenarios. Next, we calibrated the correlated effects and the errors in variables so that we obtain for the true coefficient α of 0.5 an asymptotic upward bias of 0.2 in the cross–sectional dimension (for the usual between estimator) and a downward bias of 0.2 in the time series dimension (for the usual within estimator).[26] Finally, we introduced a fair amount of x related heteroscedasticity in the model by assuming that α is not constant but randomly distributed across individuals (with mean 0.5 and standard deviation 0.2), and thus adding to the model another disturbance term of the form $(\alpha_i - \alpha)x_{it}$.[27]

For each of the three scenarios we experimented with six panels of different length and size, covering a set of values comparable to those found in many empirical studies. We combined two time spans: a short one ($T = 3$) and an average one ($T = 6$), with three cross-section sizes: a small, a medium and a large one ($N = 100$, $N = 400$, and $N = 1600$). For all eighteen resulting configurations we performed one hundred Monte-Carlo replications, on the basis of which we can compare the distributions of the different estimators.

For each simulated sample, in addition to the usual estimators (so called total, between, within, first differences and long differences), we computed four types of Chamberlain's and GMM estimators. These estimators correspond to a sequence of plausible specification errors that an econometrician, without knowing of course the true model specification, might be willing to compute and compare. The first three types of estimators match our three scenarios, being respectively based on the

[26] These values are large but remain in the plausible set. Note that, since the (asymptotic) biases of the different estimators are linear functions of these values, simulations with other values do not add much to the analysis.

[27] The relative amount of heteroscedasticity generated by this additional disturbance term may be on the low side (see Mairesse, 1990). Note that this term is an additional source of serial correlation in the disturbances, but does not affect the validity of the past and future values of x as instruments.

assumptions of NCE, CE, and CE+EV specifications. The fourth type assumes a CE+WS specification encompassing the first three ones.

These four types of Chamberlain's and GMM estimators are based on four sets of estimating equations or orthogonality conditions which are sequentially nested They can thus be directly compared in terms of increasing robustness and decreasing efficiency, and they allow for straightforward specification tests. The CE+WS specification only requires the past x to be uncorrelated with the current u_{it} which implies $(T-1)(T-2)/2$ orthogonality conditions for α, that is respectively 1 and 10 for $T=3$ and $T=6$. The CE+EV specification allows the future x to be also uncorrelated with the current u_{it}, which results in $T(T-2)$ (equals $(T-1)(T-2)/2+(T-1)(T-2)/2+(T-2)$) orthogonality conditions for α, that is respectively 3 and 24 for $T=3$ and $T=6$. In the CE specification all the x (present, past and future) are uncorrelated with u_{it} which results in $T(T-1)$ (equals $T(T-2)+T$) orthogonality conditions, that is respectively 6 and 30 for $T=3$ and $T=6$. In the NCE specification all the x's are also uncorrelated with the individual effects μ_i which leads to T^2 (equals $T(T-1)+T$) orthogonality conditions, that is respectively 9 and 36 for $T=3$ and $T=6$.

For all four assumed specifications we have computed two Chamberlain estimators: CHAM1 and CHAM2. CHAM1 is obtained using as weight matrix the inverse of the covariance matrix of π estimated under the assumption of homoscedasticity, while CHAM2 is based on the heteroscedasticity consistent estimated covariance matrix of π.[28] We have also computed the two comparable GMM estimators: GMM1 and GMM2. GMM1 is the first step estimator using as weight matrix for the orthogonality conditions the inverse of the second order moment matrix of the appropriate instruments, while GMM2 is the corresponding optimal second step estimator.

To summarize the behaviour of our different estimators, we computed the means and standard deviations of their observed distribution over the 100 Monte-Carlo replications. The discussion of the consistency and efficiency to which we turn next is mainly based on these summary statistics as given in Tables 5.2–5.5. For all simulated samples we also computed the asymptotic standard errors of the different estimators and compared their means over the Monte-Carlo replications with the observed standard deviations of the estimators (obtained from the Monte-Carlo replications).[29] These results are shown in Tables 5.6–5.8. We also performed several conditional and unconditional χ^2 specification tests (of overidentifying restrictions) for all four assumed specifications. The results are summarized in Tables 5.9 and 5.10, giving the number of rejections for the 5% significance level over the 100 replications.

---

[28] CHAM2 is theoretically optimal for the NCE and CE specifications but not fully so for the CE+EV and CE+WS specifications. For these two specifications the weight matrix $\widehat{V}_{\widehat{\pi}}^{-1}$ is not strictly the optimal one, since it does not account for the the fact that $E(\underline{x}_i\underline{x}_i')$, the variance–covariance matrix of the x's, is estimated.

[29] In fact we prefered to use the square roots of the means of the asymptotic variances rather than the means of the square roots. However, this does not seem to make any difference.

5.5.2 Consistency and Bias

The overall plausibility of our simulation design can be appreciated by comparing the usual panel data estimates on the simulated data and those which we found for real data sets and on which we largely based our calibration choices. Table 5.1 presents such estimates of the elasticity of capital parameter in a Cobb–Douglas production function (with constant return to scale) for three samples of manufacturing firms in France, Japan and the US. Table 5.2 shows, for our three scenarios and the average period—medium size configuration ($T = 6$ and $N = 400$), the means over the 100 Monte Carlo replications of the corresponding usual panel data estimates of our parameter of interest α. The results are very similar for the other five sample configurations.

The basic divergence between the cross-sectional and time series estimates, the fact that the total and between regression estimates tend to be significantly larger than the within, the first and long differences estimates, which is found in the three country samples, is reproduced in the simulated samples when the CE and CE+EV scenarios are enacted. The other revealing discrepancy among the group of time series estimates, namely the fact that the first differences estimates tend to be lower

Table 5.1 Usual panel data estimates of a Cobb-Douglas production function*
French, Japan and US Manufacturing (1967–1979)

Type of Estimator	France N = 441 T = 13	Japan N = 845 T = 13	USA N = 462 T = 13
Total	0.303	0.452	0.221
	[0.009]	[0.007]	[0.007]
	0.174	0.292	0.154
Between	0.313	0.469	0.222
	[0.031]	[0.023]	[0.024]
	0.192	0.326	0.163
Within	0.196	0.278	0.213
	[0.011]	[0.009]	[0.008]
	0.052	0.082	0.096
First differences	0.260	0.183	0.289
	[0.014]	[0.010]	[0.009]
	0.064	0.031	0.149
Long differences	0.163	0.359	0.178
	[0.039]	[0.029]	[0.030]
	0.038	0.151	0.073

In each cell, the first number is the estimated elasticity of capital, the second in brackets is the estimated standard error and the third is the regression R^2.
*$Log\,(Q_{it}/L_{it}) = \alpha Log\,(C_{it}/L_{it}) + \mu_i + u_{it}$, where Q is the deflated sales, L is the number of employees, and C is the gross book value adjusted for inflation.

Table 5.2 Usual estimates for three different (true) specifications
Means and standard deviations for 100 replications Simulated samples for N = 400 and T = 6

Type of estimator	True specification		
	NCE	CE	CEEV
Total	0.501	• 0.659	• 0.619
	[0.052]	[0.052]	[0.052]
	0.198	0.303	0.272
Between	0.501	• 0.674	• 0.673
	[0.056]	[0.057]	[0.059]
	0.211	0.333	0.323
Within	0.507	0.507	• 0.232
	[0.032]	[0.032]	[0.028]
	0.123	0.123	0.036
First differences	0.503	0.503	• 0.162
	[0.046]	[0.046]	[0.036]
	0.074	0.074	0.020
Long differences	0.503	0.503	• 0.289
	[0.047]	[0.047]	[0.050]
	0.186	0.186	0.071

In each cell the first number and second one in brackets are respectively the mean and standard deviation of the estimated α for 100 replications, while the third number is the mean of the regression R^2 for the 100 replications.
• Denotes that the estimator is inconsistent.

than the within and the long differences ones, which can clearly be seen for the Japanese sample, is also exhibited by the CE+EV scenario.[30] Note also, that the R^2 of the different regressions for the simulated samples in this last scenario are rather close to the corresponding R^2 of the regressions for the real samples.

Going beyond the usual set of estimators, we can look at the consistency of the more sophisticated Chamberlain and GMM estimators in Tables 5.3, 5.4 and 5.5. These tables give the means and standard deviations of these estimators (as computed from the Monte-Carlo replications) for the twelve combinations of assumed specifications (NCE, CE, CE+EV, CE+WS) and true specifications (NCE, CE, CE+EV). These tables are given for the three average period configurations ($T = 6$). We do not report here the results for the three other configurations ($T = 3$), but we find that the different estimators behave very similarly.

As expected, the Chamberlain and GMM estimators appear to be consistent when they should be, that is when the assumed specification is the true one or when the assumed specification encompasses the true one. On the other hand, these estimators are biased when the assumed specification is wrong, that is when it is incompatible with the true specification; these cases are marked with a bullet in the Tables.

[30] These discrepancies among the usual panel data estimates of the production function are much more pronounced when the returns to scale are not constrained to be one. See Mairesse (1990) for more details.

Table 5.3 Chamberlain's and GMM estimates under four different specification assumptions for three different (true) specifications
Means and standard deviations for 100 replications Simulated samples for N = 100 and T = 6

Assumed	Type of	True specification		
specification	estimator	NCE	CE	CEEV
NCE	CHAMB1	0.510	• 0.560	• 0.343
		[0.061]	[0.062]	[0.057]
	CHAMB2	0.498	• 0.551	• 0.334
		[0.071]	[0.075]	[0.073]
	GMM1	0.513	• 0.668	• 0.629
		[0.089]	[0.094]	[0.094]
	GMM2	0.504	• 0.606	• 0.497
		[0.069]	[0.075]	[0.078]
CE	CHAMB1	0.506	0.507	• 0.236
		[0.068]	[0.068]	[0.059]
	CHAMB2	0.497	0.497	• 0.231
		[0.077]	[0.077]	[0.074]
	GMM1	0.503	0.503	• 0.172
		[0.097]	[0.097]	[0.078]
	GMM2	0.498	0.498	• 0.205
		[0.075]	[0.075]	[0.068]
CE+EV	CHAMB1	0.510	0.510	0.415
		[0.114]	[0.115]	[0.174]
	CHAMB2	0.505	0.507	0.413
		[0.120]	[0.120]	[0.197]
	GMM1	0.500	0.495	0.275
		[0.177]	[0.178]	[0.271]
	GMM2	0.504	0.503	0.370
		[0.122]	[0.122]	[0.211]
CE+WS	CHAMB1	0.554	0.549	0.292
		[0.303]	[0.304]	[0.385]
	CHAMB2	0.555	0.545	0.311
		[0.318]	[0.317]	[0.396]
	GMM1	0.546	0.529	0.210
		[0.369]	[0.366]	[0.398]
	GMM2	0.562	0.548	0.287
		[0.312]	[0.308]	[0.398]

• Denotes that the estimator is inconsistent. In each cell the first number is the mean and the second the standard deviation of brackets are respectively the mean and standard deviation of $\hat{\alpha}$ over 100 replications.

Nevertheless, it can be seen that for the smaller samples ($N = 100$) large finite sample biases show up in the two extreme cases in which the true specification is CE+EV and the assumed ones are CE+EV and CE+WS. In the first case (CE+EV if CE+EV) the means of the CHAMB1, CHAMB2, and GMM2 estimates are around 0.40, while in the second case (CE+EV if CE+WS) they are around 0.30. The means of the GMM1 estimates are even lower about respectively 0.30 and 0.20 in these

Table 5.4 Chamberlain's and GMM estimates under four different specification assumptions for three different (true) specifications
Means and standard deviations for 100 replications Simulated samples for N = 400 and T = 6

Assumed specification	Type of estimator	True specification		
		NCE	CE	CEEV
NCE	CHAMB1	0.505	• 0.555	• 0.333
		[0.031]	[0.031]	[0.030]
	CHAMB2	0.504	• 0.555	• 0.325
		[0.031]	[0.033]	[0.031]
	GMM1	0.502	• 0.659	• 0.619
		[0.052]	[0.052]	[0.052]
	GMM2	0.504	• 0.571	• 0.409
		[0.031]	[0.034]	[0.039]
CE	CHAMB1	0.506	0.506	• 0.230
		[0.032]	[0.032]	[0.028]
	CHAMB2	0.505	0.505	• 0.229
		[0.033]	[0.033]	[0.02]
	GMM1	0.503	0.503	• 0.162
		[0.046]	[0.046]	[0.036]
	GMM2	0.505	0.505	• 0.217
		[0.032]	[0.032]	[0.029]
CE+EV	CHAMB1	0.512	0.512	0.481
		[0.059]	[0.059]	[0.098]
	CHAMB2	0.513	0.513	0.480
		[0.063]	[0.063]	[0.104]
	GMM1	0.526	0.527	0.422
		[0.091]	[0.091]	[0.143]
	GMM2	0.514	0.514	0.472
		[0.062]	[0.062]	[0.104]
CE+WS	CHAMB1	0.551	0.550	0.468
		[0.175]	[0.173]	[0.282]
	CHAMB2	0.549	0.549	0.461
		[0.182]	[0.181]	[0.289]
	GMM1	0.570	0.569	0.368
		[0.211]	[0.211]	[0.354]
	GMM2	0.551	0.550	0.453
		[0.181]	[0.180]	[0.290]

• Denotes that the estimator is inconsistent. See Table 5.3

two cases. Although these estimators, given the small sample size configuration, are extremely imprecise with computed dispersions across the 100 replications of respectively 0.2 and 0.4 in the two cases, the t tests of departure from the true value of 0.50 are quite significant.[31]

[31] If $\overline{\widehat{\alpha}}$ and $s_{\widehat{\alpha}}$ are the mean and standard deviation of the distribution of the $\widehat{\alpha}$ estimator computed from R replications, then the standard deviation of the mean estimate $\overline{\widehat{\alpha}}$ is $s_{\widehat{\alpha}}/\sqrt{R}$ and the t ratio is $\sqrt{R}(\overline{\widehat{\alpha}} - 0.5)/s_{\widehat{\alpha}}$. Here, we have for example: $t = \sqrt{100}(0.4 - 0.5)/0.2$ and $t = \sqrt{100}(0.3 - 0.5)/0.4$, which are both roughly equal to -5.

Table 5.5 Chamberlain's and GMM estimates under four different specification assumptions for three different (true) specifications
Means and standard deviations for 100 replications Simulated samples for N = 1600 and T = 6

Assumed specification	Type of estimator	True specification		
		NCE	CE	CEEV
NCE	CHAMB1	0.502	• 0.553	• 0.330
		[0.013]	[0.013]	[0.013]
	CHAMB2	0.502	• 0.553	• 0.321
		[0.014]	[0.014]	[0.013]
	GMM1	0.503	• 0.662	• 0.622
		[0.023]	[0.023]	[0.022]
	GMM2	0.502	• 0.560	• 0.387
		[0.014]	[0.015]	[0.016]
CE	CHAMB1	0.502	0.502	• 0.226
		[0.015]	[0.015]	[0.013]
	CHAMB2	0.502	0.502	• 0.223
		[0.015]	[0.015]	[0.014]
	GMM1	0.501	0.501	• 0.161
		[0.021]	[0.021]	[0.016]
	GMM2	0.502	0.502	• 0.218
		[0.015]	[0.015]	[0.014]
CE+EV	CHAMB1	0.503	0.503	0.493
		[0.027]	[0.027]	[0.049]
	CHAMB2	0.503	0.503	0.492
		[0.027]	[0.028]	[0.050]
	GMM1	0.503	0.503	0.460
		[0.050]	[0.050]	[0.076]
	GMM2	0.503	0.503	0.491
		[0.028]	[0.028]	[0.051]
CE+WS	CHAMB1	0.522	0.522	0.506
		[0.091]	[0.090]	[0.152]
	CHAMB2	0.522	0.522	0.502
		[0.093]	[0.092]	[0.149]
	GMM1	0.516	0.516	0.442
		[0.123]	[0.121]	[0.208]
	GMM2	0.522	0.521	0.501
		[0.094]	[0.092]	[0.152]

• Denotes that the estimator is inconsistent. See Table 5.3

It is reassuring to see, however, that for moderate sample sizes ($N = 400$) and even more so for larger sample sizes ($N = 1600$), the finite sample biases have largely vanished. Note though that in most cases of the CE+WS specification these biases tend to remain statistically significant, given that they decrease more slowly than the standard deviations. Note also that the behaviour of the GMM1 estimator (as computed here) is different and less satisfactory than that of the other three estimators.

5.5.3 *Efficiency and Robustness*

The simulation results are also instructive from the perspective of the efficiency of the different estimators as measured by their standard deviations computed over the replications. These numbers are given in brackets in Tables 5.6–5.8 for all the cases when the estimators are consistent, that is when the assumed specification encompasses the true specification or is the true one. We can thus compare them in various ways, not only across the different estimators (for given samples sizes and assumed specifications), but also across sample sizes, and across assumed specifications.

Looking first at the efficiency of the four estimators for given sample sizes and assumed specification, we do not find any real surprise. We can first check that CHAMB2, the (nearly) optimal Chamberlain estimator, and GMM2, the optimal GMM estimator, are practically equivalent.[32] This is an indication that taking the covariance matrix $E(\underline{x}_i\underline{x}_i')$ as known, in order to simplify the computation of the

Table 5.6 Simulated standard deviations and estimated standard errors of the GMM and Chamberlain estimators under four different specification assumptions for three different (true) specifications

Standard deviations and means of the estimated standard errors of the estimates for 100 replications
Simulated samples for N = 100 and T = 3 and 6

True specification		NSE				CE			CEEV	
Assumed specification		NSE	CE	CEEV	CEWS	CE	CEEV	CEWS	CEEV	CEWS
T = 3	CHAMB1	[0.081]	[0.111]	[1.322]	[33.24]	[0.111]	[1.351]	[5.157]	[3.161]	[9.776]
		(0.076)	(0.116)	(1.276)	(50.11)	(0.116)	(1.245)	(5.257)	(3.091)	(18.16)
	CHAMB2	[0.087]	[0.120]	[1.349]	[33.24]	[0.122]	[1.352]	[5.157]	[3.233]	[9.776]
		(0.071)	(0.110)	(1.253)	(50.11)	(0.111)	(1.223)	(5.257)	(3.074)	(18.16)
	GMM1	[0.111]	[0.128]	[1.291]	[33.24]	[0.128]	[1.340]	[5.157]	[3.144]	[9.776]
		(0.097)	(0.135)	(2.091)	(8776)	(0.135)	(2.160)	(53.83)	(46.74)	(577.0)
	GMM2	[0.089]	[0.119]	[1.437]	[33.24]	[0.120]	[1.487]	[5.157]	[7.124]	[9.776]
		(0.073)	(0.114)	(2.049)	(8776)	(0.114)	(2.067)	(53.83)	(45.35)	(577.0)
T = 6	CHAMB1	[0.061]	[0.068]	[0.114]	[0.303]	[0.068]	[0.115]	[0.304]	[0.174]	[0.385]
		(0.052)	(0.061)	(0.098)	(0.315)	(0.061)	(0.098)	(0.312)	(0.152)	(0.360)
	CHAMB2	[0.071]	[0.077]	[0.120]	[0.318]	[0.077]	[0.120]	[0.317]	[0.197]	[0.396]
		(0.040)	(0.049)	(0.084)	(0.292)	(0.049)	(0.084)	(0.291)	(0.128)	(0.339)
	GMM1	[0.089]	[0.097]	[0.177]	[0.369]	[0.097]	[0.178]	[0.366]	[0.271]	[0.398]
		(0.091)	(0.088)	(0.175)	(0.429)	(0.88)	(0.175)	(0.426)	(0.255)	(0.472)
	GMM2	[0.069]	[0.075]	[0.122]	[0.312]	[0.075]	[0.122]	[0.308]	[0.211]	[0.398]
		(0.041)	(0.052)	(0.089)	(0.320)	(0.052)	(0.089)	(0.317)	(0.140)	(0.370)

For each estimator the first number (in brackets) is the standard deviation of the estimated α over 100 replications, and the second number (in parentheses) is the mean of the estimated standard errors over the 100 replications.

[32] They are strictly identical for the assumed specification CE+WS and $T = 3$, i.e., when there is only one orthogonality condition and thus no weighting is involved.

Table 5.7 Simulated standard deviations and estimated standard errors of the GMM and Chamberlain estimators under four different specification assumptions for three different (true) specifications

Standard deviations and the mean of the estimated standard errors of the estimates for 100 replications Simulated samples for N = 400 and T = 3 and 6

True specification		NSE				CE			CEEV	
Assumed specification		NSE	CE	CEEV	CEWS	CE	CEEV	CEWS	CEEV	CEWS
T = 3	CHAMB1	[0.037]	[0.059]	[0.543]	[1.858]	[0.059]	[0.540]	[3.543]	[1.468]	[20.35]
		(0.039)	(0.059)	(0.602)	(2.497)	(0.059)	(0.598)	(5.347)	(1.854)	(70.53)
	CHAMB2	[0.038]	[0.061]	[0.549]	[1.858]	[0.060]	[0.544]	[3.543]	[1.584]	[20.35]
		(0.038)	(0.058)	(0.600)	(2.497)	(0.059)	(0.595)	(5.347)	(1.843)	(70.53)
	GMM1	[0.046]	[0.065]	[0.551]	[1.858]	[0.065]	[0.545]	[3.543]	[1.463]	[20.35]
		(0.049)	(0.068)	(0.631)	(13.06)	(0.068)	(0.625)	(92.59)	(9.708)	(6958)
	GMM2	[0.038]	[0.060]	[0.555]	[1.858]	[0.060]	[0.545]	[3.543]	[2.812]	[20.35]
		(0.038)	(0.058)	(0.627)	(13.06)	(0.058)	(0.621)	(92.59)	(9.608)	(6958)
T = 6	CHAMB1	[0.031]	[0.032]	[0.059]	[0.175]	[0.032]	[0.059]	[0.173]	[0.098]	[0.282]
		(0.028)	(0.032)	(0.053)	(0.170)	(0.032)	(0.053)	(0.170)	(0.088)	(0.267)
	CHAMB2	[0.031]	[0.033]	[0.063]	[0.182]	[0.033]	[0.063]	[0.181]	[0.104]	[0.289]
		(0.026)	(0.031)	(0.051)	(0.167)	(0.031)	(0.051)	(0.167)	(0.085)	(0.262)
	GMM1	[0.052]	[0.046]	[0.091]	[0.211]	[0.046]	[0.091]	[0.211]	[0.143]	[0.354]
		(0.046)	(0.044)	(0.090)	(0.223)	(0.044)	(0.090)	(0.223)	(0.153)	(0.367)
	GMM2	[0.031]	[0.032]	[0.062]	[0.181]	[0.032]	[0.062]	[0.180]	[0.104]	[0.290]
		(0.026)	(0.031)	(0.052)	(0.172)	(0.031)	(0.052)	(0.172)	(0.088)	(0.282)

See Table 5.6

Chamberlain estimator in the CE+EV and CE+WS cases may have no consequences in practice.[33] We can then verify that the two step estimator GMM2 is indeed more efficient than the one step estimator GMM1 by a (somewhat limited) factor of about 1.5. Finally, we can observe that the CHAMB1 estimator, which would be optimal in the case of homoscedastic errors, is in fact not less efficient than the more general CHAMB2 estimator. This should be related to the relatively modest amount of heteroscedasticity in our simulations, since the x related heteroscedasticity we have introduced, although substantial in terms of parameter heterogeneity (with $\sqrt{E(\alpha_i - \alpha)^2} = 0.2$), appears small as compared to the overall variability of the errors.[34]

Looking next at that the efficiency of our four Chamberlain and GMM estimators with sample size, we know a priori that it should increase as $\sqrt{N}$, and this is indeed verified in the simulations. In nearly all cases the standard deviations are divided by a factor of about 2, when going from $N = 100$ to $N = 400$ and

[33] Further experiments not reported here showed that significant differences between the fully optimal Chamberlain estimators and the nearly optimal ones only occur in rather peculiar cases, where we have to assume that the within component of x is much larger than the between component, and is strongly correlated with the past values of the u_{it} disturbances.

[34] However, experimenting with twice as much x heteroscedasticity also did not show up in our results.

Table 5.8 Simulated standard deviations and estimated standard errors of the GMM and Chamberlain estimators under four different specification assumptions for three different (true) specifications
Standard deviations and the means of the estimated standard errors of the estimates for 100 replications Simulated samples for N = 1600 and T = 3 and 6

True specification		NSE				CE			CEEV	
Assumed specification		NSE	CE	CEEV	CEWS	CE	CEEV	CEWS	CEEV	CEWS
T = 3	CHAMB1	[0.020]	[0.029]	[0.298]	[0.404]	[0.029]	[0.294]	[0.396]	[0.598]	[1.465]
		(0.020)	(0.029)	(0.306)	(0.393)	(0.029)	(0.304)	(0.390)	(0.601)	(1.547)
	CHAMB2	[0.021]	[0.029]	[0.296]	[0.404]	[0.029]	[0.292]	[0.396]	[0.587]	[1.465]
		(0.019)	(0.029)	(0.305)	(0.393)	(0.029)	(0.304)	(0.390)	(0.601)	(1.547)
	GMM1	[0.026]	[0.034]	[0.298]	[0.404]	[0.034]	[0.294]	[0.396]	[0.600]	[1.465]
		(0.024)	(0.034)	(0.310)	(0.406)	(0.034)	(0.308)	(0.403)	(0.813)	(6.040)
	GMM2	[0.021]	[0.029]	[0.297]	[0.404]	[0.029]	[0.293]	[0.396]	[0.601]	[1.465]
		(0.019)	(0.029)	(0.309)	(0.406)	(0.029)	(0.308)	(0.403)	(0.812)	(6.040)
T = 6	CHAMB1	[0.013]	[0.015]	[0.027]	[0.091]	[0.015]	[0.027]	[0.089]	[0.049]	[0.152]
		(0.014)	(0.016)	(0.027)	(0.088)	(0.016)	(0.027)	(0.088)	(0.047)	(0.153)
	CHAMB2	[0.014]	[0.015]	[0.027]	[0.093]	[0.015]	[0.028]	[0.091]	[0.050]	[0.149]
		(0.014)	(0.016)	(0.027)	(0.088)	(0.016)	(0.027)	(0.088)	(0.046)	(0.153)
	GMM1	[0.023]	[0.021]	[0.050]	[0.123]	[0.021]	[0.050]	[0.121]	[0.076]	[0.208]
		(0.023)	(0.022)	(0.045)	(0.110)	(0.022)	(0.045)	(0.110)	(0.079)	(0.206)
	GMM2	[0.014]	[0.015]	[0.028]	[0.093]	[0.015]	[0.028]	[0.092]	[0.051]	[0.152]
		(0.014)	(0.016)	(0.027)	(0.088)	(0.016)	(0.027)	(0.088)	(0.048)	(0.158)

See Table 5.6.

from there to $N = 1600$. Again exceptions are found in the extreme cases of the CE+WS assumed specification for the short period samples ($T = 3$) for which the standard deviations are very large (and probably not well measured with 100 replications). In contrast, the improvement in efficiency with the time dimension is not a straightforward matter: it depends on the orthogonality conditions involved and on the weight matrix used.[35] Accordingly, it should vary with the assumed specification, which we can see indeed. When moving from $T = 3$ to $T = 6$ the standard deviations of the estimators are roughly divided by a factor of 1.5, 2, and 10 for the NCE, CE and CE+EV specifications respectively. They are reduced by a factor which can be even much larger (from 10 to 100) when assuming the CE+WS specification.

Last, but specially compelling, are the findings on the trade off between efficiency and robustness. The differences of efficiency of the estimators across the assumed specifications (for given true specifications) are of similar orders of magnitudes as their observed differences of efficiency between the short and average period samples (for given assumed and true specifications). In the case of the longer samples ($T = 6$), the standard deviations are thus increased by a factor

[35] It should also be remembered that inverting the covariance matrix of the orthogonality conditions (or the covariance matrix of π) implies that T cannot be too large with regard to N.

of about 2 when going from the assumptions of NCE or CE to the weaker one of CE+EV, and increased again by a factor of about 3 when going to the even weaker assumption of CE+WS. For the shorter samples ($T = 3$) these efficiency factors are much larger, and they vary even more for the more extreme CE+WS assumption.[36] We thus have an efficiency factor between the NCE and CE specifications of roughly 1.5, and a factor between the CE and CE+EV specifications of roughly 10 (!).[37]

It is worth pointing out that contrary to one's first intuition, the loss of efficiency when choosing the CE specification rather than the much stronger NCE specification becomes negligible when $T = 6$, and remains moderate even when $T = 3$. The explanation lies in the fact that the T additional orthogonality conditions (of the NCE specification compared to the CE specification) are strongly interrelated, and that in a way they cannot do "much more" than one such condition only.[38] The crucial issue when relying on estimators based on the CE specification is not efficiency but the fact that the biases arising from other potential misspecifications such as errors in variables or simultaneity can be greatly aggravated. This is indeed what we can see, as already noted, when the true specification is CE+EV and not CE. On the other hand, efficiency becomes an important consideration when the econometrician suspects that the true specification is indeed CE+EV or CE+WS and wants to avoid such aggravated biases. In this situation efficiency can be an even more important problem than the present simulation results suggest, if the serial correlation of x is very high, and thus if the past x's are poor instruments for the changes in the current x (see Griliches and Mairesse, 1998 and Blundell and Bond, 1998 and 2000).

5.5.4 Standard Errors

Besides discussing the relative efficiency of the estimators, it is also instructive to ask whether they are efficient in absolute terms, that is whether they are precise enough to ensure a satisfactory approximation of the parameters of interest. In other words, are the standard deviations of the estimators sufficiently small? Since in practice the answer is mainly based on the estimated asymptotic standard errors of the parameter estimates, it is interesting to ascertain that these standard errors are indeed consistently estimated. This can be done by verifying that the means of the estimated

[36] As already noted, in the EV+WS case, the short sample estimates can be extremely imprecise. For $T = 3$, they are in fact based on only one orthogonality condition, and follow a Cauchy distribution with infinite variance (!).

[37] The observed efficiency factor between the CE+EV and CE+WS estimators varies very much depending on the true specification and the sample size N. It is equal to about: 1.5 if $N = 1600$ and the true specifications is NSE or CE, to about 3 if $N = 1600$ and the true specification is CE+EV, and it ranges from 3 to 25 depending on the true specification if $N = 100$ or 400.

[38] This is related to the fact that most of the variance in the x's and the y is cross-sectional ("between", not "within").

Table 5.9 General tests of specification based on the Chamberlain and GMM estimators under four different specification assumptions for three different (true) specifications
Number of rejections out of 100 replications Simulated samples for N = 100, 400 and 1600 and T = 3 and 6

	Tests		N = 100			N = 400			N = 1600		
			True specification			True specification			True specification		
			NSE	CE	CEEV	NSE	CE	CEEV	NSE	CE	CEEV
T = 3	NSE	CHAM	24	• 30	• 82	7	• 33	• 100	11	• 91	• 100
		GMM	4	• 10	• 56	7	• 28	• 100	10	• 90	• 100
	CE	CHAM	22	22	• 14	8	9	• 12	89	5	• 25
		GMM	9	11	• 1	8	7	• 9	6	5	• 25
	CEEV	CHAM	12	12	8	9	9	5	1	2	9
		GMM	8	7	4	7	7	3	1	2	8
	CEWS	CHAM	–	–	–	–	–	–	–	–	–
		GMM	–	–	–	–	–	–	–	–	–
T = 6	NSE	CHAM	84	• 90	• 99	18	• 45	• 100	6	• 96	• 100
		GMM	0	• 1	• 12	3	• 19	• 100	4	• 96	• 100
	CE	CHAM	69	67	• 80	13	12	• 64	10	10	• 98
		GMM	0	0	• 2	2	2	• 33	8	8	• 98
	CEEV	GHAM	44	44	58	10	9	18	10	10	5
		GMM	2	1	0	4	3	4	8	8	2
	CEWS	CHAM	16	14	14	7	7	8	8	8	8
		GMM	3	2	0	3	5	3	6	7	6

• Denotes the situations in which the null hypothesis is not true.

standard errors of the estimators, as computed over the replications, agree well, that is within a margin of error, with the simulated standard deviations of the estimators as obtained from the same replications. The mean standard errors are given in parentheses under the standard deviations given in brackets in Tables 14.6–14.8. When comparing these numbers, we must remember that they are themselves both estimated over replications and known with a margin of error. Making simplifying assumptions based on asymptotic normality, we can say that they are estimated independently with an absolute standard error of $\sigma/\sqrt{2R}$ where σ denotes their assumed common mean value and $R = 100$ is the number of replications, that is with a relative error of $1/\sqrt{2R}$ or about 7%. We thus can accept that they do not differ significantly at the 5% significance level, if they do not differ by more than 20% (that is approximately $(1.96\sqrt{2})$ times 7%). We shall also consider for simplicity that the standard deviations of the estimators are satisfactorily small if they are less than 0.1 in absolute value, that is if they provide an approximation of about 20% for a coefficient of 0.5.

It is reassuring to see that in most cases the estimated standard errors and the simulated standard deviations of the estimators are indeed quite close, and well within the 20% margin. Again, as could be expected, the main exceptions arise in the short samples ($T = 3$) for the estimators assuming the CE+WS specification,

Table 5.10 Nested tests of specification based on the Chamberlain and GMM estimators under four different specification assumptions for three different (true) specifications
Number of rejections out of 100 replications Simulated samples for N = 100, 400 and 1600 and T = 3 and 6

	Tests		N = 100			N = 400			N = 1600		
			True specification			True specification			True specification		
			NSE	CE	CEEV	NSE	CE	CEEV	NSE	CE	CEEV
T = 3	CEEV	CHAM	12	12	8	9	9	5	1	2	9
	CEWS	GMM	8	7	4	7	7	3	1	2	8
	CE/	CHAM	22	22	• 14	8	9	• 12	89	5	• 25
	CEWS	GMM	9	11	• 1	8	7	• 9	6	5	• 25
	CE/	CHAM	16	16	• 12	4	5	• 12	13	11	• 35
	CEEV	GMM	10	10	• 9	4	4	• 11	12	11	• 34
	NSE/	CHAM	24	• 30	• 82	7	• 33	• 100	11	• 91	• 100
	CEWS	GMM	4	• 10	• 56	7	• 28	• 100	10	• 90	• 100
	NSE/	CHAM	18	• 32	• 84	7	• 35	• 100	13	• 94	• 100
	CEEV	GMM	9	• 10	• 59	4	• 31	• 100	13	• 94	• 100
	NSE/	CHAM	13	• 26	• 83	5	• 44	• 100	7	• 97	• 100
	CE	GMM	4	• 11	• 71	5	• 40	• 100	5	• 97	• 100
T = 6	CEEV	CHAM	45	45	52	9	9	15	5	6	4
	CEWS	GMM	3	2	4	6	6	7	4	4	4
	CE/	CHAM	66	64	• 77	14	13	• 65	10	10	• 98
	CEWS	GMM	2	2	• 11	4	4	• 43	9	9	• 98
	CE/	CHAM	61	56	• 63	13	21	• 80	6	5	• 100
	CEEV	GMM	4	4	• 16	7	7	• 61	5	5	• 100
	NSE/	CHAM	85	• 92	• 100	20	• 56	• 100	8	• 96	• 100
	CEWS	GMM	2	• 2	• 2	5	• 24	• 100	4	• 96	• 100
	NSE/	CHAM	80	• 89	•100	19	• 63	• 100	2	• 98	• 100
	CEEV	GMM	1	• 7	• 34	5	• 39	• 100	0	• 98	• 100
	NSE/	CHAM	63	• 84	• 99	14	• 71	• 100	1	• 100	• 100
	CE	GMM	2	• 11	• 39	6	• 53	• 100	1	• 100	• 100

• Denotes the situations in which the null hypothesis is not true.

and thus based on only one orthogonality condition. In this case the differences are large for $N = 100$ and remain so for $N = 1600$.

It is also interesting to note that the estimated standard errors of the optimal Chamberlain and GMM estimators CHAMB2 and GMM2 tend to be significantly too optimistic for the longer and smaller samples ($T = 6$ and $N = 100$) and the NSE, CE and CE+EV assumed specifications. This supports the findings by Arellano and Bond (1991), who caution against the possible underestimation of the standard errors of the optimal GMM estimator.[39]

[39] Some further simulations seem to indicate that such finite sample under-estimation occurs when the number of orthogonality conditions is large (or not small enough) compared to the size (N) of the sample. When this happens, the GMM and Chamberlain estimators of the α parameter tend also to suffer from sizeable finite sample bias (compared to the bias of the OLS estimator). See Bound, Jaeger and Baker (1993).

If we focus now on the absolute magnitude of the standard deviations of the estimators, we get a clearcut picture, highlighting again the important trade off between robustness and efficiency. The precision is acceptable for the estimators assuming the NCE and CE specifications. By contrast, for the estimators assuming the CE+WS specification, it is just acceptable only for the longer and larger samples ($T = 6$ and $N = 1600$) and if the true specification is NCE or CE (but not CE+EV). For the estimators assuming the CE+EV specification, the situation is haldway, their precision becoming acceptable for the longer samples ($T = 6$), with the exception of the case of the CE+EV true specification with $N = 100$.

5.5.5 Specification Tests

An important question that we also want to illustrate with our simulations, is the performance of the specification tests used to decide whether the specification assumed by the econometrician can be accepted or should be rejected. These are the χ^2 tests of overidentifying restrictions (presented in Sect. 5.3.3), which we have implemented in the two different ways corresponding to the optimal Chamberlain and GMM estimators. The Chamberlain type tests (denoted as CHAM) are based on the whole set of $\left(T^2\right)$ estimating equations, thus involving the N_s parameters of secondary interest in addition to the parameter α of primary interest, and they use the $\left(T^2 \times T^2\right)$ CHAM2 weight matrix (i.e., the inverse of the heteroscedasticity consistent Π covariance matrix estimate).[40] The GMM type tests (denoted as GMM) are based on a reduced set of $\left(T^2 - N_s\right)$ orthogonality conditions which are derived from the full set of $\left(T^2\right)$ estimating equations (or orthogonality conditions) by eliminating the N_s parameters of secondary interest, and they thus only involve here the unique parameter α of primary interest. They use the $\left(\left(T^2 - N_s\right) \times \left(T^2 - N_s\right)\right)$ GMM2 weight matrix (i.e., the inverse of the orthogonality conditions covariance matrix first step estimate).[41]

We consider both the general tests of the four assumed specifications (NCE), (CE), (CE+EV) and (CE+WS), and the corresponding difference or nested tests of one specification conditional on another (more general) one, that is (CE+EV) given (CE+WS), (CE) given (CE+EV) or (CE+WS), (NCE) given (CE) or (CE+EV) or (CE+WS).[42] Tables 5.9 and 5.10 report on these general and difference tests, by

[40] N_s is equal to 0, T, $2T$ and $((T(T+1)/2 + (T-1))$ respectively for the NCE, CE, CE+EV and CE+WS assumed specifications.

[41] While the CHAM tests correspond to Wald type tests of the restriction on Π, the GMM tests can be viewed as Lagrange multiplier type tests. See also following footnotes 42 and 44.

[42] The χ^2 of the nested tests are simply computed as the differences of the χ^2 of the corresponding general tests. Our GMM tests are thus implemented with a different weight matrix for the null and alternative hypotheses, while the CHAM tests are computed holding the weight matrix constant for both hypotheses. It is usually considered to be better in practice (in finite samples) to hold the weight matrix constant (as estimated under the null hypothesis, or the alternative hypothesis or even a more general hypothesis). Note that the general specification tests can be simply viewed as difference tests conditional on the alternative "specification" of the unrestricted Π matrix.

giving the number of rejections obtained out of the 100 replications with a 5% significance level. They do it for our eighteen configurations of true specifications and sample sizes. The number of rejections is an estimate of the true size of the test (for a 5% nominal size) when the tested specification is true (i.e., when it encompasses the underlying true specification or is identical to it), and an estimate of the power of the test when the tested specification is false. With 100 replications only, these estimates of the tail probabilities of the tests statistics distribution cannot be very accurate and they should be taken with some caution. They are nonetheless quite suggestive.[43]

The most striking observation is that the CHAM tests tend to reject much more frequently than the GMM tests in the small and medium size samples ($N = 100$ and $N = 400$). This could be expected, though perhaps not to such extent, since one would expect that, loosely speaking, the CHAM weight matrices be "larger" than the GMM ones (or the Π covariance matrix estimated with no restrictions be "smaller" than the ones estimated with restrictions).[44] For the large samples ($N = 1600$) the two types of tests give quite similar results.

Another way to point out the same finding is to note that the true size of the CHAM tests tends to be much higher in the smaller samples than the (5%) nominal size (i.e., they overreject when the tested specification is true), while the performance of the GMM tests is more satisfactory in this respect. However, and conversely, the power of the CHAM tests tend to be greater than that of the GMM tests in the smaller samples. If we take the example of the test of the (CE) specification in the longer period medium size sample ($T = 6, N = 400$), the percentages of rejection are respectively about 12% and 2% for the CHAM and GMM tests, when the (CE) specification is correct (i.e., if the underlying true specification is (NCE) or (CE)), and about 64% and 33% when it is not (i.e., if the underlying true specification is CE+EV). Note also in Table 5.10 that for this same example the power of both tests goes up to respectively 80% and 61% when the (CE+EV) is rightly assumed as the alternative hypothesis (instead of implicitly assuming the unrestricted Π matrix specification).[45]

If we want to summarize, the practical conclusion is that in small (and not so small) samples the weighting can matter much for the specification tests, while it seems to matter only little for the estimation of the parameters of interest. This could also be expected, since (loosely speaking again) what matters for the tests is the absolute magnitude of the weights, and for the estimation their relative magnitude. Of course more important than the weighting, and the choice between imple-

[43] We have checked, however, in a number of cases, that the percentage of rejections do in fact change very little when performing 1000 replications instead of 100.

[44] It is known that when testing linear restrictions in the linear regression model, the χ^2 statistics are larger for the Wald tests than for the Lagrange multiplier tests, and hence that the former reject more frequently the null hypothesis than the latter for a given significance level. See Berndt and Savin (1977) and Breusch (1979).

[45] The corresponding percentages of rejection when we perform 1000 replications instead of 100 are: 16 and 4 (instead of 12 and 2), 57 and 30 (instead of 64 and 33), and 71 and 57 (instead of 80 and 61).

menting the Chamberlain method, or GMM in this respect, remains the validity and relevance of the instruments, that is their exogeneity and their correlation with the regressors.

5.6 Appendix A: An Extended View of the Chamberlain Method

In this Appendix we show how the generalization of the Chamberlain method presented in 5.2.5 applies to simultaneous equations models, vector autoregressions (VAR) and endogenous attrition models.

5.6.1 Simultaneous Equations Models

Simultaneous equations models on panel data fit straightforwardly within the general Chamberlain framework. Using the same notations as before, but considering that we now have J dependent variables y with $\underline{y}_i' = \left(y_{i1}^{(1)}, \ldots, y_{iT}^{(1)}, y_{i1}^{(2)} \ldots, y_{iT}^{(J)}\right)$, these models can be writen as:

$$A_{11}(\underline{\theta})\underline{y}_i + A_{12}(\underline{\theta})\underline{x}_i = \underline{v}_i.$$

If we assume that no restrictions are made on the covariance matrices $E(\underline{x}_i\underline{x}_i')$ and $E(\underline{v}_i\underline{v}_i')$ of the explanatory variables x and of the disturbances v, and applying the result on the elimination of parameters and equations of Sect. 5.2.5, we need only to focus on the moment conditions relative to the joint covariance matrix $E(\underline{y}_i\underline{x}_i')$ of the $\underline{y}$'s and $\underline{x}$'s. Assuming also for simplicity that the $\underline{x}$'s are strictly exogenous, (i.e., $E(\underline{v}_i\underline{x}_i') = 0$), and that the model is not underidentified (the matrices A_{11} and A_{12} having appropriate structures), the relevant estimating equations for an efficient estimation of the parameters of interest $\underline{\theta}$ are the following:

$$A_{11}(\underline{\theta})E(\underline{y}_i\underline{x}_i') + A_{12}(\underline{\theta})E(\underline{x}_i\underline{x}_i') = 0$$
or
$$A_{11}(\underline{\theta})\Pi + A_{12}(\underline{\theta}) = 0.$$

5.6.2 VAR Models

Panel data Vector Autoregressive (VAR) models can also be considered in the Chamberlain framework. The distinction between dependent and explanatory variables is no longer relevant, and we have a set of endogenous variables explained by their past values. To keep notations simple, let us take the case of a stationary VAR

model of order one for two variables x_i and y_i. We can write this model as:

$$\begin{cases} y_{it} = \varphi_{yy} y_{it-1} + \varphi_{yx} x_{it-1} + d_t^y + \mu_i^y + u_{it}^y \\ x_{it} = \varphi_{xy} y_{it-1} + \varphi_{xx} x_{it-1} + d_t^x + \mu_i^x + u_{it}^x \end{cases} \quad t = 1, \ldots, T$$

where d_t^y and d_t^x are year constants; μ_i^y and μ_i^x are individual effects; u_{it}^y and u_{it}^x are year varying disturbances.[46] As in the case of the lagged dependent variable model of Sect. 5.2.3.4, we can also rewrite it in matrix form as

$$(I - \Phi \otimes L) \begin{pmatrix} \underline{y}_i \\ \underline{x}_i \end{pmatrix} = \begin{pmatrix} \underline{d}^y \\ \underline{d}^x \end{pmatrix} + \begin{pmatrix} \underline{v}_i^y \\ \underline{v}_i^x \end{pmatrix}$$

with

$$\begin{pmatrix} \underline{v}_i^y \\ \underline{v}_i^x \end{pmatrix} = \zeta \otimes \underline{l}_1 + \mu \otimes \underline{l} + \begin{pmatrix} \underline{u}_i^y \\ \underline{u}_i^x \end{pmatrix}$$

where L is the lag matrix, $\Phi = \begin{bmatrix} \varphi_{yy} & \varphi_{yx} \\ \varphi_{xy} & \varphi_{xx} \end{bmatrix}$, $\zeta = \begin{pmatrix} \zeta_i^y \\ \zeta_i^x \end{pmatrix} = \Phi \begin{pmatrix} y_{i0} \\ x_{i0} \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_i^y \\ \mu_i^x \end{pmatrix}$, and $\underline{l}_1' = (1, 0, \ldots, 0)$. We are thus led to the same general formulations as (5.21) or (5.22), by setting the $A(\theta)$ matrix equal to $(I - \Phi \otimes L)$, with variables centered at their year means. With both y and x being endogenous, the model is not identified and restrictions have to be imposed for identification on the covariance matrix $\Omega(\underline{\theta}) = E(\underline{v}_i \underline{v}_i')$ of the disturbances.

The usual identification assumption in VAR models is that the time varying errors u_{it}^x and u_{it}^y are not serially correlated (but may be contemporaneously correlated). The covariance matrix of the disturbances can then be written as:

$$\Omega(\underline{\theta}) = \begin{bmatrix} V_{\underline{u}_i^y} & E(\underline{u}_i^y \underline{u}_i^{x\prime}) \\ E(\underline{u}_i^x \underline{u}_i^{y\prime}) & V_{\underline{u}_i^x} \end{bmatrix} + V_\zeta \otimes \underline{l}_1 \underline{l}_1' + V_\mu \otimes \underline{l}\underline{l}' + E(\zeta \mu') \otimes \underline{l}_1 \underline{l}' + E(\mu \zeta') \otimes \underline{l}\underline{l}_1'$$

where $V_{\underline{u}_i^y}$, $V_{\underline{u}_i^x}$ and $E(\underline{u}_i^x \underline{u}_i^{y\prime})$ are the $(T \times T)$ diagonal matrices of the time varying variances of the disturbances $\underline{u}_i^x$, and $\underline{u}_i^{y\prime}$ and of their covariances; V_μ and V_ζ are the two (2×2) symmetric covariance matrices of μ_i^x and μ_i^y and ζ_i^x and ζ_i^y, and $E(\zeta \mu')$ is the (2×2) matrix of their covariances.

For a total of $2T(2T+1)/2 = T(2T+1)$ estimating equations derived from (5.22), we have thus $(3T+14)$ parameters, that is four parameters of primary interest in the Φ of lagged coefficients (φ_{yy} φ_{yx} φ_{xy} and φ_{yy}), and $3T+10$ parameters of secondary interest in the covariance matrix Ω of the disturbances. These estimating equations are nonlinear (in the φ parameters), and the direct implementation of ALS could be quite complicated.

---

[46] An interesting non-stationary VAR model for panel data is proposed in Chamberlain (1982) and in Holtz-Eakin, Newey and Rosen (1988), in which the error terms are of the form $\Psi_t \mu_i + u_{it}$ (i.e., with interactions between the individual effects μ_i and time effects Ψ_t), and in which the coefficients $\varphi_{yy}, \varphi_{yx}, \varphi_{xy}$ and φ_{xx} matrix may also be varying over time.

There is actually a much better way to solve the problem. It is easy to see that for each time period s, both y_{is} and x_{is} can be expressed as functions of the current and past values of the idiosyncratic disturbances $u^y_{is}, \ldots u^y_{i1}$ and $u^x_{is}, \ldots, u^x_{i1}$ and the individual effects μ^y_i and μ^x_i, as well as the unknown initial observations y_{i0} and x_{i0}. Therefore, under the maintained assumption of no serial correlation in the time varying disturbances, their correlation with the future values of the residuals is only a simple function of the 10 parameters of V_ζ, V_μ and $E(\zeta'\mu)$. And thus the residuals $(v^y_{it} - v^y_{it-1})$ and $(v^x_{it} - v^x_{it-1})$ of the model in first differences are not correlated with the past values of the y and x variables lagged by two periods and more (i.e., $y_{it-2}, \ldots, y_{i1}$ and $x_{it-2}, \ldots, x_{i1}$), and a subset of the estimating equations is more simply:

$$E\left(\begin{pmatrix}\Delta v^y_{it}\\ \Delta v^x_{it}\end{pmatrix}(y_{is}, x_{is})\right) = 0 \quad \forall s < t-1.$$

This second set of estimating equations, derived in Holtz-Eakin, Newey and Rosen (1988), has the advantage of eliminating all the parameters of secondary interest and that of being linear in the parameters of primary interest. However, it provides only $2(T-1)(T-2)$ estimating equations, and there are $2T(T+1) - 2(T-1)(T-2) = 8T-4$ equations lost for only $3T+10$ parameters eliminated. It follows that the corresponding estimator is not the most efficient. The efficient set of estimating equations is derived in Ahn and Schmidt (1995).

Arellano and Bover (1995) and Blundell and Bond (1998) show that assuming stationarity adds additional estimating equations that can substantially improve efficiency. Indeed, stationarity implies that the initial values y_{i0} and x_{i0} are themselves functions of the parameters of primary interest, the individual effects and the infinite number of idiosyncratic disturbances prior to the beginning of the sample, so their covariances with the idiosyncratic disturbances for the period of observation can be expressed as functions of the parameters of primary interest only.

5.6.3 Endogenous Attrition

The general Chamberlain framework can also be helpful in dealing with endogenous attrition. For example, Abowd et al. (1995) consider the case of the simple first order autoregressive model $y_{it} = \rho y_{it-1} + \mu_i + u_{it}$, which can also be written in vector form as:

$$[I - \rho L]\underline{y}_i = \mu_i \underline{l} + \rho y_{i0}\underline{l}_1 + \underline{u}_i = \underline{v}_i.$$

The individuals i are firms, appearing and disappearing at individual dates b_i and d_i. The authors consider the general attrition processes compatible with what Rubin (1976) calls data missing at random, where the probability of dying at date d given the firm history $y^{(d)}_i = (y_{ib_i}, \ldots, y_{id_i})$ is equal to the probability given the latent variable $y^*_i = (y_{ib_i}, \ldots, y_{iT})$, that is:

$$P\left(d_i = d|y_i^{(d)}\right) = P\left(d_i = d|y_i^*\right). \tag{5.46}$$

In this case, for firms appearing at a same date b, we can write

$$\begin{aligned}\Omega(\theta) &= E\left(\underset{\sim}{v}_i\underset{\sim}{v}_i'\right) = [I-\rho L]\,E\left(\underset{\sim}{y}_i^*\underset{\sim}{y}_i^{*\prime}\right)[I-\rho L]' \qquad (5.47)\\ &= [I-\rho L]\sum_{d=1}^{T} P(d_i=d)\,E\left(\underset{\sim}{y}_i^*\underset{\sim}{y}_i^{*\prime}|d_i=d\right)[I-\rho L]'\\ &= [I-\rho L]\sum_{d=1}^{T} P(d_i=d)\,E\left(E\left(\underset{\sim}{y}_i^*\underset{\sim}{y}_i^{*\prime}|y_i^{(d)}\right)|d_i=d\right)[I-\rho L]'\\ &= [I-\rho L]\sum_{d=1}^{T} P(d_i=d)\,E\left(f_d\left(y_i^{(d)},\theta\right)|d_i=d\right)[I-\rho L]'\end{aligned}$$

where the probability density of dying at time d is a function $f_d\left(y_i^{(d)},\theta\right)$ of the firm history $y_i^{(d)}$ and the unknown parameters θ.

Assuming normality, the conditional expectation $E[f_d(y_i^{(d)},\theta)|d_i=d]$ can be written as $g_d[E(y_i^{(d)}|d_i=d), E(y_i^{(d)}y_i^{(d)\prime}|d_i=d),\theta]$ showing that (5.47) can be taken as estimating equations for the parameters θ and the set of additional auxiliary parameters $\{E(y_i^{(d)}|d_i=d)$ and $E(y_i^{(d)}y_i^{(d)\prime}|d_i=d)$, for $d=1,\ldots,T\}$.

5.7 Appendix B: Vector Representation of the Chamberlain Estimating Equations

We show here how to write in a vector format the Chamberlain estimating equations written in matrix form. After recalling the main properties of the Vec operator, we use it to obtain the vector representation of the Chamberlain estimating equations in the cases of the CE, EV and WS specifications and their combinations, as well as in the cases of a lagged dependent variable specification and of the existence of restrictions on the covariance matrix of the disturbances.

5.7.1 The Vec Operator

The Vec. operator transform a matrix into a vector by stacking one above the other the different columns of the matrix. This operator has many properties, and we only give here a few important ones (see Magnus and Neudecker, 1988, for a general presentation).

If A, B and C are three matrices of conforming size, we can write:

$$\mathrm{Vec}(ABC) = \left(C' \otimes A\right) \mathrm{Vec}(B)$$

which also implies that

$$\mathrm{Vec}(AB) = \left(I_{\mathrm{ncol}(B)} \otimes A\right) \mathrm{Vec}(B) \quad \text{and} \quad \mathrm{Vec}(BC) = \left(C' \otimes I_{\mathrm{nrow}(B)}\right) \mathrm{Vec}(B).$$

If $V_1$and V_2 are two vectors of any size, we have also

$$\mathrm{Vec}(V_1 V_2') = \mathrm{Vec}(V_2 \otimes V_1)\,.$$

and for two matrices A and B with the same number of columns, we have:

$$\mathrm{Vec}(A)'\mathrm{Vec}(B) = Tr(A'B)\,.$$

5.7.2 Correlated Effects

The estimating equations (see (5.12)) are in matrix form the following:

$$\underline{\Pi} = \underline{b}' \otimes I_T + \underline{l}\underline{c}'\,.$$

Taking the Vec operator after transposing the matrices leads to

$$\underline{\pi} = \mathrm{Vec}(\Pi') = \mathrm{Vec}(\underline{b} \otimes I_T) + (\underline{l} \otimes I_{KT})\,\underline{c} = \mathrm{Vec}(\underline{b} \otimes I_T) + G_{ce}\underline{c}\,.$$

Writing $\underline{b} = \sum_{k=1}^{K} b_k \underline{l}_k$, where $(\underline{l}_1, \ldots, \underline{l}_K)$ is the canonical base of R^K, we obtain:

$$\begin{aligned}\mathrm{Vec}(\underline{b} \otimes I_T) &= \mathrm{Vec}\left(\textstyle\sum_{k=1}^{K} b_k \underline{l}_k \otimes I_T\right) = \textstyle\sum_{k=1}^{K} b_k \mathrm{Vec}(\underline{l}_k \otimes I_T) \\ &= \left[\mathrm{Vec}(\underline{l}_1 \otimes I_T) | \cdots | \mathrm{Vec}(\underline{l}_K \otimes I_T)\right] \underline{b}\end{aligned}$$

The estimating equations can thus be written in vector form as:

$$\underline{\pi} = [G_0 | G_{ce}] \frac{\underline{b}}{\underline{c}} = H_{ce}\underline{\gamma}.$$

which is a most simple case since the matrix H is constant with only coefficients equal to zero or equal to 1.

5.7.3 Errors in Variables

The estimating equations (see (5.14)) are in this case:

$$\underline{\Pi} = \underline{b}' \otimes I_T \left[I - V_e E(\underline{x}_i \underline{x}_i')^{-1}\right]\,.$$

They depend on the specific parameterization of the covariance matrix V_e of the measurement errors on the variables x.

Let us consider here the usual case where these errors are assumed to be serially uncorrelated, but where they can be correlated with each other. V_e is thus the following $KT \times KT$ matrix $V_e = \left[\Sigma_{k,l}\right]_{k,l=1,\ldots,K}$ where $\Sigma_{k,l}$ is the covariance matrix of the two $T \times 1$ vectors of the measurement errors for the variables k and l. These errors being non serially correlated, the matrices $\Sigma_{k,l}$ are diagonal and can be written as: $\Sigma_{k,l} = \text{Diag}\,(\sigma_{k,l,1}, \ldots, \sigma_{k,l,T})$, leading to:

$$\begin{aligned}\underline{\pi} &= \text{Vec}(\Pi') = \text{Vec}(\underline{b} \otimes I_T) - \text{Vec}\left(E(\underline{x}_i \underline{x}_i')^{-1} V_e \left[\underline{b} \otimes I_T\right]\right) \\ &= G_0 \underline{b} - \left(I_T \otimes E(\underline{x}_i \underline{x}_i')^{-1}\right) \text{Vec}\left(V_e \left[\underline{b} \otimes I_T\right]\right)\end{aligned}$$

In this expression, the product $V_e \left[\underline{b} \otimes I_T\right]$ has the form:

$$\begin{aligned}V_e \left[\underline{b} \otimes I_T\right] &= \begin{bmatrix} \sum_{l=1}^K b_l \Sigma_{1,l} \\ \vdots \\ \sum_{l=1}^K b_l \Sigma_{K,l} \end{bmatrix} = \begin{bmatrix} \text{Diag}_{t=1,\ldots,T}\left(\sum_{l=1}^K b_l \sigma_{1,l,t}\right) \\ \vdots \\ \text{Diag}_{t=1,\ldots,T}\left(\sum_{l=1}^K b_l \sigma_{1,l,t}\right) \end{bmatrix} \\ &= \begin{bmatrix} Diag_{t=1,\ldots,T}(\zeta_{1,t}) \\ \vdots \\ \text{Diag}_{t=1,\ldots,T}(\zeta_{K,t}) \end{bmatrix} = \sum_{k=1}^K \sum_{t=1}^T \underline{l}_k \otimes D_t \zeta_{k,t} \\ &= F_{ev} \underline{\zeta}\end{aligned}$$

where $\underline{\zeta}' = (\zeta_{11}, \ldots, \zeta_{1T}, \ldots, \zeta_{K1}, \ldots, \zeta_{KT})$ with $\zeta_{k,t} = \sum_{l=1}^K b_l \sigma_{k,l,t}$, and where $\underline{l}_k$ is the kth element of the canonical base of R^K, and D_t is the matrix with all zero coefficients except the t^{th} one of the diagonal which is equal to 1. Note that we cannot identify all parameters, since the coefficients $\sigma_{k,l,t}$ are only entering $V_e \left[\underline{b} \otimes I_T\right]$ via $\zeta_{k,t} = \sum_{l=k}^K b_l \sigma_{k,l,t}$, and only these functions of the $\sigma_{k,l,t}$ are identifiable.

We can thus write in vector form:

$$\text{Vec}(V_e \left[\underline{b} \otimes I_T\right]) = \sum \text{Vec}(\underline{l}_k \otimes D_t) \zeta_{k,t} = M_{ev} \underline{\zeta} \qquad k = 1, \ldots, K \; t = 1, \ldots, T$$

and therefore also the estimating equations as:

$$\begin{aligned}\underline{\pi} &= G_0 \underline{b} - \left[I_T \otimes E(\underline{x}_i \underline{x}_i')^{-1}\right] M_{ev} \underline{\zeta} \\ &\left[G_0 | G_{ev}\left(E(\underline{x}_i \underline{x}_i')\right)\right] \begin{pmatrix} \underline{b} \\ \underline{\zeta} \end{pmatrix} = H_{ev}(E(\underline{x}_i \underline{x}_i'))\underline{\gamma}.\end{aligned}$$

The expression of the covariance matrix, as given in (5.36) involves the gradient $\partial H \underline{\theta} / \partial \underline{m}'$, with $\underline{m}$ being here= $\text{Vec}E(\underline{x}_i \underline{x}_i')$. Using the relation $d\text{Vec}(A^{-1}) = -A'^{-1} \otimes A^{-1} d\text{Vec}(A)$ (see Magnus and Neudecker, 1988), we can compute this gradient as:

$$\begin{aligned}\partial H\underline{\theta}^0/\partial\underline{m}' &= -\left(\Phi(\beta)\otimes I_{KT}\right)\left(E(\underline{x}_i\underline{x}_i')^{-1}\otimes E(\underline{x}_i\underline{x}_i')^{-1}\right)\\ &= -(\Phi(\beta)E(\underline{x}_i\underline{x}_i')^{-1})\otimes E(\underline{x}_i\underline{x}_i')^{-1}\\ &= -(\Pi^0 - b^{0\prime}\otimes I_T)\otimes E(\underline{x}_i\underline{x}_i')^{-1}\ .\end{aligned}$$

5.7.4 Weak Simultaneity

The case of weak simultaneity is quite similar to that of errors in variables. The relation between the parameters of interest and the auxiliary parameters includes the second order moment matrix of the explanatory variables. Here the $T\times KT$ matrix $\Phi(\underline{\beta})$ is formed of blocks of $T\times T$ upper triangular matrices, and thus can be written as:

$$\Phi(\underline{\beta}) = \sum_{k=1}^{K}\sum_{i\le j} Mij(k)\beta_{i,j,k}\,,$$

where the matrices $M_{i,j}(k)$ have zero coefficients except in the k th $T\times T$ block where the (i,j) coefficients are equal to one. We thus have:

$$\mathrm{Vec}\left(\Phi\left(\underline{\beta}\right)'\right) = \sum_{k=1}^{K}\sum_{i\le j}\mathrm{Vec}\left(Mij(k)'\right)\underline{\beta}_{i,j,k} = M_{ws}\underline{\beta}\ .$$

and the estimating equations can be rewritten as:

$$\begin{aligned}\underline{\pi} &= G_0 b - \left(I_T\otimes E(\underline{x}_i\underline{x}_i')^{-1}\right)M_{ws}\underline{\zeta}\\ &= \left[G_0|G_{ws}\left(E(\underline{x}_i\underline{x}_i')\right)\right]\begin{pmatrix}\underline{b}\\ \underline{\zeta}\end{pmatrix} = H_{ws}(E(\underline{x}_i\underline{x}_i'))\underline{\gamma}.\end{aligned}$$

5.7.5 Combination of the Different Cases

The estimating equations in vector form, when combining the previous specification errors, are easily obtained by juxtaposition of the matrices G_0, G_{ce}, $G_{ev}(E(\underline{x}_i\underline{x}_i'))$ and $G_{ws}(E(\underline{x}_i\underline{x}_i'))$ in the preceding formulas. For example, in the case of both correlated effects and errors in variables (EC+EV), we can write:

$$\underline{\pi} = \left[G_0|G_{ev}\left(E(\underline{x}_i\underline{x}_i')\right)|G_{ce}\right]\begin{pmatrix}\underline{b}\\ \underline{\zeta}\\ \underline{c}\end{pmatrix} = H_{ce,ev}(E(\underline{x}_i\underline{x}_i'))\underline{\gamma}.$$

5.7.6 Lagged Dependent Variable

In the dynamic case, the estimating equations take the form:

$$[I - \underline{\alpha} L]\,\Pi = \underline{b}' \otimes I_T + \underline{l}_1 \underline{\lambda}' + \underline{l}\,\underline{\delta}\,.$$

Applying the Vec operator after transposition of the matrices, we have:

$$\begin{aligned}\underline{\pi} &= [(L \otimes I_{KT})\,\underline{\pi}]\,\underline{\alpha} + G_0 \underline{b} + \underline{l}_1 \otimes I_{KT}\,\underline{\lambda} + \underline{l} \otimes I_{KT}\,\underline{\delta} \\ &= [(L \otimes I_{KT})\,\underline{\pi} | G_0 | \underline{l}_1 \otimes I_{KT} | \underline{l} \otimes I_{KT}] \begin{pmatrix} \underline{\alpha} \\ \underline{b} \\ \underline{\lambda} \\ \underline{\delta} \end{pmatrix} = H_{\mathrm{LDV}}(\underline{\pi})\gamma.\end{aligned}$$

5.7.7 Restrictions on the Covariance Matrix of the Disturbances

Let us give a last example in the cases of restrictions on the covariance matrix of the disturbances (see Sect. 5.2.4). In such cases, when using the Vec. operator to transform the estimating equations, one has to be careful to avoid the duplications due to the symmetry of variance matrices. The vector transformation of a symmetric matrix which picks up only the different elements of a matrix is referred as the Vech operator. One can switch from one operator to the other by premultiplication of a given matrix. Considering for example a symmetric matrix M of size L, one can define two matrices D_L and D_L^+ of size $(L(L+1)/2) \times L^2$ and $L^2 \times (L(L+1)/2)$ respectively, such that $\mathrm{Vech}(M) = D_L\,\mathrm{Vec}(M)$ and $\mathrm{Vec}(M) = D_L^+ \mathrm{Vech}(M)$.

In the frequent cases where the covariance matrix of the disturbances Σ is linear in a parameter vector $\underline{\lambda}$, it is possible using the Vech operator to write simply:

$$\mathrm{Vech}\,(\Sigma\,(\underline{\lambda})) = K\underline{\lambda}\,.$$

Taking the most standard case of the error components model, we have $\Sigma(\underline{\lambda}) = \sigma_B^2 J + \sigma_W^2 I$, where σ_B^2 and σ_W^2 are the respective variances of the individual effects μ_i and the idiosyncratic disturbances u_{it}, I is the unity matrix and J the matrix all the coefficients of which are equal to 1. We can write:

$$\mathrm{Vech}\left(\Sigma\left(\sigma_B^2, \sigma_W^2\right)\right) = [\mathrm{Vech}\,(J)\,|\mathrm{Vech}\,(I)] \begin{pmatrix} \sigma_B^2 \\ \sigma_W^2 \end{pmatrix},$$

or without assuming the constancy over time of the variances of the idiosyncratic disturbances:

$$\text{Vech}\left(\Sigma\left(\sigma_B^2, \sigma_{W_1}^2, \ldots, \sigma_{W_T}^2\right)\right) = \left[\text{Vech}\left(J\right) | \text{Vech}\left(M_1\right) | \ldots | \text{Vech}\left(M_T\right)\right] \begin{pmatrix} \sigma_B^2 \\ \sigma_{W_1}^2 \\ \vdots \\ \sigma_{W_T}^2 \end{pmatrix}$$

where the M_t matrices have all their coefficients equal to zero except the t th diagonal coefficient equal to 1.

5.8 Appendix C: Manipulation of Equations and Parameters in the ALS Framework

ALS is a flexible method allowing the manipulation of equations and parameters without loosing asymptotic efficiency. Provided some simple rules are observed, one can change the way in which the estimating equations are written in order to eliminate auxiliary parameters as well as some parameters of secondary interest, while preserving the asymptotic properties of the estimators.

5.8.1 Transformation of the Estimating Equations

The estimating equations can be transformed without efficiency loss in the estimation of the parameter of interest for a wide range of transformations. This is summarized in the following result.

Let f be a function of the three arguments $\underline{\theta}, \underline{\mu}$, and g where $g \in R^{n_g}$ with values in R^{n_g}, such that: $f(\underline{\theta}, \underline{\mu}, g) = 0 \Leftrightarrow g = 0$ and $(\frac{\partial f}{\partial g}(\underline{\theta}^0, \underline{\mu}^0, 0)$ is invertible. Then the optimal ALS estimator based on $h(\underline{\theta}^0, \underline{\mu}^0) = f(\underline{\theta}^0, \underline{\mu}^0, g(\underline{\theta}^0, \underline{\mu}^0)) = 0$ has the same asymptotic properties as the optimal ALS estimator based on $g(\underline{\theta}^0, \underline{\mu}^0) = 0$.

This result has several implications. When the number of estimating equations is equal to the dimension of the auxiliary parameter, these equations can be transformed so that they take the form of a direct restriction $\underline{\mu}^0 - h\left(\underline{\theta}^0\right) = 0$. As already explained, this has several advantages in the implementation of the optimal ALS estimator (essentially that it does not require a two steps procedure).

Also, important computational simplifications arise when the estimating equations can be rewritten as a linear function of the parameter of interest such as $\underline{\pi}^0 - H\left(\underline{m}^0\right)\underline{\theta}^0 = 0$. Let us consider again the case of the autoregressive model. We have seen in Sect. 5.2.3 that the estimating equations can be written linearly in terms of the parameters of primary interest $\underline{\alpha}$ and $\underline{b}$, and secondary interest $\underline{\delta}$ and $\underline{\lambda}$, as

$$[I - \alpha L]\Pi = M(\underline{b})I_T + \underline{l}\underline{\delta}' + \underline{l}_1\underline{\lambda}'.$$

If we premultiply them by $[I-\alpha L]^{-1}$, we transform them in the form of a direct restriction on π, but which is highly nonlinear in terms of the parameters of interest. The two sets of equations are equivalent, but the first is more convenient to deal with and provides an explicit expression for the estimator of the parameter of interest.

5.8.2 Eliminating Parameters of Secondary Interest

The parameters of interest $\underline{\theta}$ are often divided into a subset of parameters of primary interest and a subparameter set of secondary interest.[47] It may be convenient to only estimate the first set, specially when they have a small dimension and the parameters of secondary interest a high one. As already mentioned, this is possible by simply eliminating the parameter of secondary interest in the estimating equations. Crépon, Kramarz and Trognon (1998) show that the potential asymptotic efficiency of the ALS estimator for the parameters of primary interest is unaffected, as long as the number of estimating equations discarded in the elimination process is equal to the number of the parameters of secondary interest.

More precisely, let $\underline{\theta}_p$ and $\underline{\theta}_s$ be the parameters of primary and secondary interest of dimension $n_{\underline{\theta}_p}$ and $n_{\underline{\theta}_s}$ and let g_p and g_s be a partition of the estimating equations of dimension n_{g_p} and n_{g_s}. Assume that $n_{g_s} = n_{\underline{\theta}_s}$, and that the $\left(n_{g_s} \times n_{\underline{\theta}_s}\right)$ square matrix $\frac{\partial g_s}{\partial \underline{\theta}_s'}(\underline{\theta}^0, \underline{\mu}^0)$ is invertible, then there exists a neigborhood N of $(\underline{\theta}^0, \underline{\mu}^0)$ and a function φ of $\underline{\theta}_p$ and $\underline{\mu}$ such that for $(\underline{\theta}, \underline{\mu})$ in N, $g_p\left(\underline{\theta}_p, \underline{\theta}_s, \underline{\mu}\right) = 0$ is equivalent to $\underline{\theta}_s = \varphi\left(\underline{\theta}_p, \underline{\mu}\right)$. If $\widehat{\underline{\theta}} = \left(\widehat{\underline{\theta}}_p, \widehat{\underline{\theta}}_s\right)$ is the optimal ALS estimator based on the full set of estimating equations

$$g\left(\underline{\theta}^0, \underline{\mu}^0\right) = 0 \tag{5.48}$$

and if $\widetilde{\underline{\theta}}_p$ is the optimal ALS estimator based on the restricted set of estimating equations

$$h(\underline{\theta}_p^0, \underline{\mu}^0) = g_p(\varphi(\underline{\theta}_p^0, \underline{\mu}^0), \underline{\theta}_p^0, \underline{\mu}^0) = 0, \tag{5.49}$$

then $\widetilde{\underline{\theta}}_p$ is asymptotically equivalent to $\widehat{\underline{\theta}}_p$.

Taking again the example of the autoregressive model, the $2KT$ parameters $\underline{\delta}$ and $\underline{\lambda}$ can be simply eliminated from the estimating equations by first differencing them and then eliminating the first year estimating equation, i.e., by premultiplying

[47] We prefer to speak of parameters of secondary interest rather than calling them nuisance parameters, since these parameters may have an interpretation.

them successively by an appropriate $(T-1)\times T$ matrix Δ and an appropriate $(T-2)\times(T-1)$ matrix E_1.[48]

Instead of using the KT^2 estimating equations to estimate the $2KT+K+1$ parameters, we use only KT^2-2KT transformed equations to estimate the $K+1$ parameters of primary interest without any loss of asymptotic efficiency.

The specification tests of Sect. 5.3.3 can also be implemented either on the reduced set of estimating (5.49) or on the whole set (5.48) including all the parameters. Under the null hypothesis of correct specification, the corresponding statistics are asymptotically equivalent. To see this consider the test statistics of (5.36), based on the implementation of the optimal estimator. They are equivalent to the statistics in (5.39) testing that the residual function of the auxiliary parameters is null once all the parameters of interest have been eliminated (see end of Sect. 5.3.3). But in both cases, the elimination can be chosen to end up with the same function of the auxiliary parameters, so the test statitics are equivalent. They follow a χ^2 distribution with degrees of freedom equal to the difference between the number of equations used in the estimation and the number of estimated parameters.[49]

5.8.3 *Recovering Parameters of Secondary Interest Once Eliminated*

Once the parameters of primary interest estimated, it is possible to obtain an estimation of the parameters of secondary interest, potentially, as efficient as if this parameter had been estimated directly. This is obtained by the solution of the following minimisation problem:

$$\widetilde{\underline{\theta}}_s = \underset{\underline{\theta}_s}{\text{Arg min}}\{g'(\underline{\theta}_s,\widetilde{\underline{\theta}}_p,\widehat{\underline{\mu}})S^* g(\underline{\theta}_s,\widetilde{\underline{\theta}}_p,\widehat{\underline{\mu}})\}, \tag{5.50}$$

with $S^* = W^{-1} = \left[\frac{\partial g^0}{\partial \underline{\mu}'} V_{\underline{\mu}} \frac{\partial g^{0\prime}}{\partial \underline{\mu}}\right]^{-1}$. Note that a simple replacement of $\underline{\theta}_p$ by $\widetilde{\underline{\theta}}_p$ in the equations $\underline{\theta}_s = \varphi(\underline{\theta}_p,\underline{\mu})$ used to eliminate the parameter of primary interest provides a consistent, but not efficient estimator $\underline{\theta}_s^{\#} = \varphi(\widetilde{\underline{\theta}}_p,\widehat{\underline{\mu}})$ of $\underline{\theta}_s$.

Let us consider the intermediate situation, in which the estimating equations take the form

---

[48] The $(T-1)\times T$ matrix Δ is defined as: $\Delta = \begin{pmatrix} -1 & 1 & & 0 \\ & & \ddots & \\ 0 & & -1 & 1 \end{pmatrix}$. It transformes any $T\times K$ matrix m with $m' = (l_1',\ldots,l_T')$ into the $(T-1)\times K$ matrix $m_\Delta = \Delta m$ with $m_\Delta' = (l_2'-l_1',\ldots,l_T'-l_{T-1}')$. The $(T-2)\times(T-1)$ E_1 matrix is simply defined as $E_1 = \begin{bmatrix} 0 & 1 & & \\ & & \ddots & \\ & & 0 & 1 \end{bmatrix}$.

[49] This is another way to understand that this quantity has to be constant along the different efficient estimation procedures that can be implemented: to keep efficiency the implicit restriction imposed on the auxiliary parameters must remain the same.

$$\begin{pmatrix} g_s(\underline{\theta},\underline{\mu}) \\ g_p(\underline{\theta},\underline{\mu}) \end{pmatrix} = \begin{pmatrix} \varphi(\underline{\theta}_p,\underline{\mu}) - \underline{\theta}_s \\ h(\underline{\theta}_p,\underline{\mu}) \end{pmatrix}$$

The efficient estimator $\widetilde{\underline{\theta}}_s$ of the parameters of secondary interest solving (5.50) is given by[50]

$$\widetilde{\underline{\theta}}_s = \varphi(\widetilde{\underline{\theta}}_p,\widehat{\underline{\mu}}) + W_{12}W_{22}^{-1}h(\widetilde{\underline{\theta}}_p,\widehat{\underline{\mu}}). \tag{5.51}$$

This last equation clearly shows that the optimal estimator of $\underline{\theta}_s$ is generally different from the estimator obtained by simple replacement $\underline{\theta}_s^{\#}$.[51]

The covariance matrix of the estimators is given by the usual formula of (5.36). It leads in this specific case to the expression

$$\begin{aligned} V(\widetilde{\underline{\theta}}_p) &= \left[\frac{\partial h'}{\partial \underline{\theta}_p} W_{22}^{-1} \frac{\partial h}{\partial \underline{\theta}_p'}\right]^{-1} \\ \text{Cov}\,(\widetilde{\underline{\theta}}_p,\widetilde{\underline{\theta}}_s) &= V(\widetilde{\underline{\theta}}_p)\left[\frac{\partial \varphi'}{\partial \underline{\theta}_2} + \frac{\partial h'}{\partial \underline{\theta}_2} W_{22}^{-1} W_{21}\right] \\ V(\widetilde{\underline{\theta}}_s) &= W_{11} - W_{12}W_{22}^{-1}W_{21} + \text{Cov}\,(\widetilde{\underline{\theta}}_p,\widetilde{\underline{\theta}}_s)' V(\widetilde{\underline{\theta}}_p)^{-1}\,\text{Cov}\,(\widetilde{\underline{\theta}}_p,\widetilde{\underline{\theta}}_s) \end{aligned} \tag{5.52}$$

Let us give two examples in which recovering parameters of secondary interest can be interesting.

1. Specification testing
Consider the case in which the set of estimating equations can be divided into two subsets

$$g' = (g_1', g_2') \in F(R^{n_{\underline{\theta}}} \to R^{n_{g_1}} \times R^{n_{g_2}}).$$

Assume we want to test for the compatibity of the first subset of estimating equations g_1 with the second one. This can be done simply in the following way: introduce an extra parameter $\underline{\delta}_1$ of dimension n_{g_1}, and consider the new ALS problem given by the estimating equations

$$\begin{aligned} g_1(\underline{\theta}^0,\underline{\mu}^0) - \underline{\delta}_1^0 &= 0 \\ g_2(\underline{\theta}^0,\underline{\mu}^0) &= 0 \end{aligned}$$

The compatibility of g_1 with g_2 can be reformulated as the assumption $\underline{\delta}_1 = 0$. So a test statistic can be computed as

$$\zeta_{12} = N\widehat{\underline{\delta}}_1' \left(\widehat{V}_{\widehat{\underline{\delta}}_1}\right)^{-1} \widehat{\underline{\delta}}_1.$$

[50] The solution of the previous problem is given by: $\widetilde{\underline{\theta}}_s = \varphi(\widetilde{\underline{\theta}}_p,\widehat{\underline{\mu}}) - (S_{11})^{-1}S_{12}h(\widetilde{\underline{\theta}}_p,\widehat{\underline{\mu}})$, and we have the relation $S_{11}^{-1}S_{12} = -W_{12}W_{22}^{-1}$ from the formula of block inversion (see Magnus and Neudecker 1988)

[51] Note, however, that when the residuals $h(\widetilde{\underline{\theta}}_2,\widehat{\underline{\mu}})$ are zero, i.e., when the parameters of primary interest are just identified, we have $\widetilde{\underline{\theta}}_1 = \underline{\theta}_1^{\#}$. Another case when these estimators are the same is when the matrix $W_{12} = 0$.

It is possible to use different procedures to estimate the parameters of the model. First eliminate $\underline{\delta}_1$ from the set of estimating equations and so estimate $\underline{\theta}$ using $g_2(\underline{\theta}, \underline{\mu}) = 0$. This provides a "robust" estimator $\widetilde{\underline{\theta}}_2$ of $\underline{\theta}_2$. It is then straightforward to implement the test. This requires only to recover an estimator $\widetilde{\underline{\delta}}_1$ of $\underline{\delta}_1$ using (5.51), and the asymptotic covariance matrix using (5.52). Note that it is possible to test individually the compatibility with the initial set of estimating equations g_2 of each equation in g_1. This means simply to test that a specific component, $\underline{\delta}_1 = 0$. Note also, that once the appropriate set of estimating equations has been selected, it is straightforward to recover an estimate of $\underline{\theta}_2$ as efficient as if it had been directly estimated on the whole set of selected estimating equations. This is simply done through another ALS step in which $\widetilde{\underline{\theta}}_2$ and $\widetilde{\underline{\theta}}_1$ are estimators of auxiliary parameters $\underline{\mu}^0 = (\underline{\mu}_1^0, \underline{\mu}_2^0) = (\underline{\theta}_2^0, \underline{\delta}_1^0)$ to be constrained by

$$\begin{pmatrix} \underline{\mu}_1^0 \\ \underline{\mu}_2^0 \end{pmatrix} \begin{pmatrix} H & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \underline{\Psi}_1^0 \\ \underline{\theta}_2^0 \end{pmatrix}$$

where H selects in g_1 the estimating equations to be removed for the estimation of $\underline{\theta}_2$, and $\underline{\Psi}_1$ the subset of parameters in $\underline{\delta}_1$ not constrained to zero.

2. The autoregressive model

The previous device to eliminate and recover some parameters of secondary interest can help to avoid nonlinear estimation. Consider the case of the autoregressive model and the case where the correlation between disturbances is restricted. As described in Sect. 5.2.4 this introduces a subset of nonlinear estimating equations. Depending on the assumptions about the covariance between the explanatory variables and the disturbances, the usual set of estimating equations, restricting the $\underline{\pi}$ matrix is of the form

$$[I - \alpha L]\,\underline{\pi} = M(\underline{b}) I_T + \Psi(\underline{\delta}) \tag{5.53}$$

or of the form

$$[I - \alpha L]\,\underline{\pi} = M(\underline{b}) I_T + \Phi(\underline{\beta}) E(\underline{x}_i \underline{x}_i')^{-1}.$$

Let us assume it is of the form of (5.53). When the correlation between the disturbances is restricted, i.e., the covariance matrix of $\underline{v}_i$ can be written as a function of the parameters $\underline{\lambda}$ of dimension less than $T(T+1)/2$: $E(\underline{v}_i \underline{v}_i') = \Sigma(\underline{\lambda})$, some additional, but nonlinear, estimating equations are available.

$$[I - \alpha L] V_w [I - \alpha L]' + \Psi(\underline{\delta}) E(\underline{x}_i \underline{x}_i') \Psi(\underline{\delta})' = \Sigma(\underline{\lambda}). \tag{5.54}$$

The direct introduction of these equations has several disadvantages discussed in Sect. 5.2.4. A way to avoid this is to proceed in several steps. The covariance matrix is first let totally free, hence, it is parameterized by $T(T+1)/2$, parameters say $\underline{\sigma}$

$$E(\underline{v}_i \underline{v}_i') = \Sigma(\underline{\sigma}) = \sum_{k \le l} \sigma_{kl} M_{kl},$$

where M_{kl} is the matrix which is zero except the elements (k,l) and (l,k) which are one.

Thus we consider the set of estimating equations formed by (5.53) and

$$[I-\alpha L]V_w[I-\alpha L]' + \Psi(\underline{\delta})E(\underline{x}_i\underline{x}_i')\Psi(\underline{\delta})' = \Sigma(\underline{\sigma}) \tag{5.55}$$

instead of (5.53) and (5.54), the only difference being that now $E(\underline{v}_i\underline{v}_i') = \Sigma(\underline{\sigma})$, instead of $\Sigma(\underline{\lambda})$.

The parameter $\underline{\sigma}$ can be eliminated from the set of estimating equations by simply excluding the second order equations (5.55). A "robust to serial correlation" and efficient estimator $\widetilde{\underline{\theta}} = (\widetilde{\underline{\alpha}}, \widetilde{\underline{\beta}}, \widetilde{\underline{\delta}})$ of the parameter of primary interest $\underline{\theta}^0 = (\underline{\alpha}^0, \underline{\beta}^0, \underline{\delta}^0)$ can be obtained on the basis of the usual (5.53) restricting the Π matrix. Following the previous methods, an efficient estimator $\widetilde{\underline{\sigma}}$ of the parameter of secondary interest $\underline{\sigma}$ can be obtained using (5.54). Now, as in the previous example, we can use the total set of parameters $(\widetilde{\underline{\sigma}}, \widetilde{\underline{\theta}})$ as an estimator of an auxiliary parameter $\underline{\mu}^0 = (\underline{\sigma}^0, \underline{\theta}^0)$ that can be further constrained to incorporate restrictions on the covariance matrix.

The main advantage of this procedure is that, for a given set of estimating (5.53) restricting the Π matrix, the estimation of the covariance matrix of the disturbances is unrestricted. Thus, if the true covariance matrix has a distinctive pattern as that coming from the error components model, we could in principle recognize it on its estimate $\widetilde{\underline{\sigma}}$. Notice that the specification test corresponding to the last ALS step is a test of the restriction implied on the covariance matrix of the disturbances. Finally, notice that the parameters of primary interest $\underline{\theta}$ can be efficiently eliminated for this last step, all the attention being devoted to the restrictions on $\underline{\sigma}$. Once these restrictions are imposed, (5.53) gives a simple way to recover an efficient ultimate estimation of $\underline{\theta}$.

5.8.4 Elimination of Auxiliary Parameters

Auxiliary parameters can be eliminated when estimating equations do not restrict them. Consider the case in which the auxiliary parameters can be divided into two sets of subparameters: $\underline{\mu}' = (\underline{\mu}_1', \underline{\mu}_2')$ and assume that μ_2 enters only a subset g_2 of the estimating equations $g' = (g_1', g_2')$ that does not restrict it, i.e., the parameters of interest $\underline{\theta}$ can be divided into $\underline{\theta}' = \left(\underline{\theta}_1', \underline{\theta}_2'\right)$ and estimating equations take the form: $g_1(\mu_1, \underline{\theta}_1) = 0, g_2(\mu_1, \mu_2, \underline{\theta}_2) = 0$ and $\partial g_2/\partial \underline{\theta}_2'$ invertible (this implies that $\dim(g_2) = \dim(\theta_2)$). Then the subset of estimating equations g_2 can be dropped for the optimal ALS estimation of the parameters of interest $\underline{\theta}_1$. This has the interesting consequence that the auxiliary parameters do not need to be estimated.

A useful application arises in the Chamberlain approach when the analysis takes into account a subset of explanatory variables but does not make any inference about it. Consider the case in which the set of explanatory variables x can be divided into two subsets $x' = [x_1', x_2']$. We can decompose the Π matrix into two

parts corresponding to the two subsets of coefficients of the linear projection of the dependent variable on the whole set of explanatory variables $\Pi = [\Pi_1, \Pi_2]$. Consider a linear model as in Sect. 5.2.2. We can write $\Pi = M(\underline{b}) + \Psi$, with by definition $\Psi = E^*(\underline{v}_i | \underline{x}_{i1}, \underline{x}_{i2})$ and decompose each matrix $M(\underline{b}), \Psi$ in two parts: $M(\underline{b}) = [M(\underline{b}_1), M(\underline{b}_2)]$ and $\Psi = [\Psi_1, \Psi_2]$. These relations do not represent a model by themselves without a specific parameterization of the matrix Ψ (or $\Phi = \Psi E(\underline{x}_i \underline{x}_i')^{-1}$) and an explicit form for $M(\underline{b})$. If we do not make any inference about the variable x_2, the matrix Π_2 is unrestricted and the related equations can be dropped. We are then left with the following equations: $\Pi_1 = M(\underline{b}_1) + \Psi_1$, or $\Pi_1 = M(\underline{b}_1) + \Phi_1 \left[E(\underline{x}_{i1} \underline{x}_{i1}') - E(\underline{x}_{i1} \underline{x}_{i2}') E(\underline{x}_{i2} \underline{x}_{i2}')^{-1} E(\underline{x}_{i2} \underline{x}_{i1}') \right]^{-1}$, which lead to a specific model with a parametrization of either Ψ_1 (i.e., $E^*(\underline{v}_i | \underline{x}_{i1} - E^*(\underline{x}_{i1} | \underline{x}_{i2}))$ or Φ_1 (i.e., $E(\underline{v}_i \underline{x}_{i1}')$).

5.9 Appendix D: Equivalence Between Chamberlain's, GMM and Usual Panel Data Estimators

The Random Effects and Correlated Effects models are respectively defined on the basis of the following panel data equation

$$y_{it} = \sum x_{it}^{(k)} b_k + \mu_i + u_{it} \qquad k = 1, \ldots, K$$

by assuming that $x_{it}^{(k)}$ is uncorrelated with both disturbances μ_i and u_{is}, or by assuming only that $x_{it}^{(k)}$ is uncorrelated with the idiosyncratic disturbance u_{is}. The most usual and well known estimators are the Generalized Least Squares (GLS) for the Random Effects model, and the Within and the First Difference estimators for the Correlated Effects model.

The GLS estimator $\widehat{\underline{b}}_{\mathrm{RE}}$ is computed using a consistent estimator $\widehat{\Omega}$ of the covariance matrix Ω of the overall disturbance $v_{it} = \mu_i + u_{is}$

$$\widehat{\underline{b}}_{\mathrm{RE}} = \left(\overline{x_i' \widehat{\Omega}^{-1} x_i} \right)^{-1} \overline{x_i' \widehat{\Omega}^{-1} \underline{y}_i}$$

where $x_i = \left[\underline{x}_i^{(1)} | \cdots | \underline{x}_i^{(K)} \right]$. For every sequence of matrices $\widehat{\Omega}$ converging in probability to Ω, we can approximately write:

$$\sqrt{N} \left(\widehat{\underline{b}}_{\mathrm{RE}} - \underline{b} \right) = \left(\overline{x_i' \Omega^{-1} x_i} \right)^{-1} \sqrt{N} \left(\overline{x_i' \Omega^{-1} \underline{v}_i} \right) + o_p(1) .$$

The within estimator $\widehat{\underline{b}}_{\mathrm{CEW}}$ is simply computed as

$$\widehat{\underline{b}}_{\mathrm{CEW}} = \left(\overline{x_i' W x_i} \right)^{-1} \overline{x_i' W \underline{y}_i},$$

where $W = I - J/T$.

The first difference estimator $\widehat{\underline{b}}_{\text{CE}\Delta}$ is computed as

$$\widehat{\underline{b}}_{\text{CE}\Delta} = \left(\overline{(\Delta x_i)' \widehat{V(\Delta \underline{v}_i)}^{-1} (\Delta x_i)} \right)^{-1} \overline{(\Delta x_i)' \widehat{V(\Delta \underline{v}_i)}^{-1} \left(\Delta \underline{y}_i \right)}.$$

for any consistent estimator $\widehat{V(\Delta \underline{v}_i)}$ of $V(\Delta \underline{v}_i)$, and we can approximately write:

$$\sqrt{N}(\widehat{\underline{b}}_{\text{CE}\Delta} - \underline{b}) = \left(\overline{(\Delta x_i)' V(\Delta \underline{v}_i)^{-1} (\Delta x_i)} \right)^{-1} \sqrt{N} \overline{(\Delta x_i)' V(\Delta \underline{v}_i)^{-1} (\Delta \underline{v}_i)} + o_p(1)$$

Our purpose here is to show that these usual estimators are special cases of the more general Chamberlain estimators. Since we know that the Chamberlain method and GMM are equivalent when based on the same second order moment restrictions implied by the model, it is enough to to show that they are special cases of the GMM estimator.

For both the Random Effects and Correlated Effects models, there is a matrix M, such that the orthogonality conditions take the specific form

$$E(\underline{X}_i \otimes (M \underline{v}_i)) = 0.$$

where $\underline{X}_i = \text{Vec}(x_i)$. In the case of Random Effects M is the $(T \times T)$ identity matrix I_T, while in the Correlated Effects case, M is the $((T-1) \times T)$ difference matrix Δ.[52]

The GMM estimator is the GLS estimator based on the orthogonality conditions

$$\left(\overline{\underline{X}_i \otimes M \underline{y}_i} \right) = \overline{\underline{X}_i \otimes (M \underline{x}_i)} b + \overline{\underline{X}_i \otimes (M \underline{v}_i)}.$$

It is thus defined as

$$\widehat{b}_{\text{GMM}} = \begin{array}{l} \left[\overline{\underline{X}_i' \otimes (M \underline{x}_i)'} W^{-1} \overline{\underline{X}_i \otimes (M \underline{x}_i)} \right]^{-1} \\ \left[\overline{\underline{X}_i' \otimes (M \underline{x}_i)'} W^{-1} \overline{\underline{X}_i \otimes \left(M \underline{y}_i \right)} \right] \end{array}$$

where $W = E\left(\underline{X}_i \otimes (M \underline{v}_i) \underline{X}_i' \otimes (M \underline{v}_i)' \right)$, and it is such that:

$$\sqrt{N}\left(\widehat{b}_{\text{GMM}} - b \right) = \begin{array}{l} \left[\overline{\underline{X}_i' \otimes (M \underline{x}_i)'} W^{-1} \overline{\underline{X}_i \otimes (M \underline{x}_i)} \right]^{-1} \\ \left[\overline{\underline{X}_i' \otimes (M \underline{x}_i)'} W^{-1} \sqrt{N} \left(\overline{X_i \otimes (M v_i)} \right) \right]. \end{array}$$

Assuming that the disturbances are homoscedastic with respect to the explanatory variables, that is:

$$E\left(\underline{v}_i \underline{v}_i' | \underline{X}_i \right) = E\left(\underline{v}_i \underline{v}_i' \right) = \Omega,$$

we can write more simply:

[52] See the definition of the difference matrix Δ in the footnote 48 in Sect. 5.8.2.

$$W = E\left(\underline{X}_i \otimes (M\underline{v}_i)\underline{X}_i' \otimes (M\underline{v}_i)'\right) = E\left(\underline{X}_i\underline{X}_i'\right) \otimes \Sigma\,,$$

with $\Sigma = ME\left(\underline{v}_i\underline{v}_i'\right)M' = M\Omega M'$. Denoting by $\underline{e}_1$ and $\underline{e}_2$ two random vectors, we can also write:

$$\begin{aligned}
\overline{\underline{X}_i' \otimes \underline{e}_{i1}'} W^{-1} \overline{\underline{X}_i \otimes \underline{e}_{i2}} &= \mathrm{Vec}\left(\overline{\underline{e}_{i1}\underline{X}_i'}\right)' E\left(\underline{X}_i\underline{X}_i'\right)^{-1} \otimes \Sigma^{-1}\mathrm{Vec}\left(\overline{\underline{e}_{i2}\underline{X}_i'}\right) \\
&= \mathrm{Vec}\left(E\left(\underline{e}_{i1}\underline{X}_i'\right) + o_p(1)\right)' \mathrm{Vec}\left(\Sigma^{-1}\overline{\underline{e}_{i2}\underline{X}_i'} E\left(\underline{X}_i\underline{X}_i'\right)^{-1}\right) \\
&= \mathrm{tr}\left(E\left(\underline{e}_{i1}\underline{X}_i'\right)' \Sigma^{-1}\overline{\underline{e}_{i2}\underline{X}_i'} E\left(\underline{X}_i\underline{X}_i'\right)^{-1}\right) + o_p(1) \\
&= \mathrm{tr}\left(\overline{\underline{X}_i' E\left(\underline{X}_i\underline{X}_i'\right)^{-1} E\left(\underline{e}_{i1}\underline{X}_i'\right)' \Sigma^{-1}\underline{e}_{i2}}\right) + o_p(1)
\end{aligned}$$

where $E\left(\underline{X}_i\underline{X}_i'\right)^{-1} E\left(\underline{e}_{i1}\underline{X}_i'\right)'$ is the vector of the coefficients of the linear projection of $\underline{e}_{i1}$ on $\underline{X}_i$.If $\underline{e}_{i1}$ is in the subspace generated by the x's, we thus have $\underline{X}_i' E\left(\underline{X}_i\underline{X}_i'\right)^{-1} E\left(\underline{e}_{i1}\underline{X}_i'\right)' = \underline{e}_{i1}'$, which leads to:

$$\overline{\underline{X}_i' \otimes \underline{e}_{i1}'} W^{-1} \overline{\underline{X}_i \otimes \underline{e}_{i2}} = \overline{\underline{e}_{i1}'\Sigma^{-1}\underline{e}_{i2}} + o_p(1)\,.$$

It follows that the element (l,k) of the matrix $\left[\overline{\underline{X}_i' \otimes (M\underline{x}_i)'} W^{-1} \overline{\underline{X}_i \otimes (M\underline{x}_i)}\right]$

$$\overline{\underline{X}_i' \otimes \left(M\underline{x}_i^{(l)}\right)'} W^{-1} \overline{\underline{X}_i \otimes \left(M\underline{x}_i^{(k)}\right)},$$

is equals to:

$$\overline{\left(M\underline{x}_i^{(l)}\right)' \Sigma^{-1}\left(M\underline{x}_i^{(k)}\right)} + o_p(1)\,.$$

Similarly, we have: [53]

$$\overline{\underline{X}_i' \otimes (M\underline{x}_i)'} W^{-1}\sqrt{N}\left(\overline{\underline{X}_i \otimes (M\underline{v}_i)}\right) = \sqrt{N}\left(\overline{x_i'M'\Sigma^{-1}Mv_i}\right) + o_p(1)\,.$$

And we can thus write that:

$$\begin{aligned}
\sqrt{N}\left(\widehat{\underline{b}}_{\mathrm{GMM}} - \underline{b}\right) &= \left[\overline{\underline{X}_i' \otimes (M\underline{x}_i)'} W^{-1} \overline{\underline{X}_i \otimes (M\underline{x}_i)}\right]^{-1} \\
&\quad \left[\overline{\underline{X}_i' \otimes (M\underline{x}_i)'} W^{-1}\sqrt{N}\left(\overline{\underline{X}_i \otimes (M\underline{v}_i)}\right)\right] \\
&= \left(\overline{x_i'M'\Sigma^{-1}(Mx_i)} + o_p(1)\right)^{-1}\sqrt{N}\left(\overline{x_i'M'\Sigma^{-1}Mv_i} + o_p(1)\right) \\
&= \left(\overline{x_i'M'\Sigma^{-1}Mx_i}\right)^{-1}\sqrt{N}\left(\overline{x_i'M'\Sigma^{-1}Mv_i}\right) + o_p(1)
\end{aligned}$$

[53] Using the fact that $E(M\underline{v}_i) = 0$ and $V(M|\underline{v}_i)$ exists, and hence that $\sqrt{N}M\underline{v}_i$ is bounded in probability and $\sqrt{N}M\underline{v}_i = o_p(1)$.

Assuming now that M is the identity matrix and $\Sigma = \Omega$ like in the Random Effects model case, we have:

$$\begin{aligned}\sqrt{N}\left(\widehat{\underline{b}}_{\text{GMM}} - \underline{b}\right) &= \left(\overline{x_i'\Omega^{-1}x_i}\right)^{-1}\sqrt{N}\overline{x_i'\Omega^{-1}v_i} + o_p(1) \\ &= \sqrt{N}\left(\widehat{\underline{b}}_{\text{RE}} - \underline{b}\right) + o_p(1)\ ,\end{aligned}$$

showing that indeed in the case of homoscedasticity, the GLS estimator, the GMM and the Chamberlain estimators are asymptotically equivalent.

In the case of the Correlated Effects model, M is the difference matrix, and we have:

$$\sqrt{N}\left(\widehat{\underline{b}}_{\text{GMM}} - \underline{b}\right) = \left(\overline{(\Delta x_i)'V(\Delta \underline{v}_i)^{-1}(\Delta x_i)}\right)^{-1}\sqrt{N}\overline{(\Delta x_i)'V(\Delta \underline{v}_i)^{-1}(\Delta v_i)} + o_p(1)$$

showing that the GMM and the Chamberlain estimators are asymptotically equivalent to the First Difference estimator.

If, moreover, the disturbances u_{it} are not serially correlated, that is if the covariance matrix $\Omega = \sigma_B^2 J + \sigma_W^2 I$, we have $\Sigma = M\Omega M' = \sigma_W^2 \Delta\Delta'$, and $M'\Sigma^{-1}M = \sigma_W^{-2}\Delta'(\Delta\Delta')^{-1}\Delta$. Noting that we have also $\Delta'(\Delta\Delta')^{-1}\Delta = W$, it follows that:[54]

$$\sqrt{N}\left(\widehat{b}_{\text{GMM}} - b\right) = \left(\overline{x_i'W_T x_i}\right)^{-1}\sqrt{N}\overline{x_i'W_T v_i} + o_p(1) = \sqrt{N}\left(\widehat{b}_W - b\right) + o_p(1)\ ,$$

showing that in the case of Correlated Effects model, the GMM and Chamberlain estimators are also asymptotically equivalent to the Within estimator when $\underline{u}_i$ is homoscedastic with respect to x and not serially correlated.

5.10 Appendix E: Design of Simulation Experiments

5.10.1 Generating Process of the Variable x

We generate the explanatory variable x as the sum of a between (or cross–sectional or permanent) component ξ_i and a within (or time series or time varying) component ξ_{it}. We suppose that the between component ξ_i is itself the sum of an exogenous subcomponent ξ_i^e and a correlated one ξ_i^c (i.e., correlated with the individual effect μ_i in the regression model). We also assume that the within component ξ_{it} is the sum of a true part ξ_{it}^* (serially correlated) and an error of measurement ξ_{it}^f serially uncorrelated. These different components are generated by four normal and independent random variables η_i, μ_i, η_{it} and φ_{it} with zero means and standard deviations σ_{η_i}, σ_{μ_i}, $\sigma_{\eta_{it}}$, $\sigma_{\varphi_{it}}$. We can thus write:

[54] $\Delta'(\Delta\Delta')^{-1}\Delta$ is the projector on the subspace generated by the columns of Δ'. Given that $\Delta = \Delta I_T = \Delta(B_T + W_T) = \Delta W_T$, this subspace is included in $Im(W_t)$. Because both subspaces are of dimension $(T-1)$, they are identical and thus $\Delta'(\Delta\Delta')^{-1}\Delta = W$.

$$
\begin{aligned}
x_{it} &= \xi_i + \xi_{it} = (\xi_i^e + \xi_i^c) + \left(\xi_{it}^* + \xi_{it}^f\right) \\
\xi_i^e &= \eta_i, \qquad \xi_i^c = \lambda \mu_i \\
\xi_{it}^* &= \textstyle\sum_{\tau=0}^{l} \omega_\tau \eta_{it-\tau}, \qquad \xi_{it}^f = \nu \varphi_{it} \,.
\end{aligned}
$$

where the weights ω_τ are such that ξ_{it}^* follows an autoregressive process of parameter ρ truncated after l years, and such that $\sum_{\tau=0}^{l} \omega_\tau^2 = 1$, implying that the variance of ξ_{it}^* and η_{it} are equal, i.e., $\mathrm{V}(\xi_{it}^*) = \mathrm{V}(\eta_{it})$. These weights are thus equal to:

$$
\omega_\tau = \rho^\tau \sqrt{\frac{1-\rho^2}{1-\rho^{2(l+1)}}} \,.
$$

5.10.2 Regression Model

We consider the true regression model:

$$
y_{it} = \alpha x_{it}^* + (\mu_i + u_{it}) \,,
$$

where x_{it}^* is the true (unobserved) value of x (i.e., $x_{it}^* = \xi_i + \xi_{it}^*$), α is the parameter of interest, and u_{it} is another independent random normal variable of mean zero and standard deviation $\sigma_{u_{it}}$. The estimated regression can thus be written as

$$
y_{it} = \alpha x_{it} + (\mu_i + (u_{it} - \alpha \nu \varphi_{it})).
$$

The dependent y_{it} can also be defined as the sum of a between component ζ_i and a within component ζ_{it}:

$$
y_{it} = \zeta_i + \zeta_{it} \,,
$$

with

$$
\begin{aligned}
\zeta_i &= \alpha \xi_i + \mu_i = \alpha(\eta_i + \lambda \mu_i) + \mu_i \\
\zeta_{it} &= \alpha \xi_{it}^* + u_{it} = \alpha \sum_{\tau=0}^{l} \omega_\tau \eta_{it-\tau} + u_{it} \,.
\end{aligned}
$$

The variances and covariances of the between and within components of the x and y variables have the following form:

$$
\begin{aligned}
\mathrm{V}(\xi_i) &= \sigma_{\eta_i}^2 + \lambda^2 \sigma_{\mu_i}^2 \\
\mathrm{Cov}\,(\zeta_i, \xi_i) &= \alpha \mathrm{V}(\xi_i) + \lambda \sigma_{\mu_i}^2 \\
\mathrm{V}\,(\zeta_i) &= \alpha^2 \sigma_{\eta_i}^2 + (1 + \alpha\lambda)^2 \sigma_{\mu_i}^2 = \alpha^2 \mathrm{V}\,(\xi_i) + (1 + 2\alpha\lambda)\, \sigma_{\mu_i}^2 \\
\mathrm{V}(\xi_{it}) &= \omega_t^2 \sigma_{\eta_i}^2 + \nu^2 \sigma_{\varphi_{it}}^2 \\
\mathrm{Cov}\,(\zeta_{it}, \xi_{it}) &= \alpha \omega_t^2 \sigma_{\eta_i}^2 = \alpha \mathrm{V}(\xi_{it}) - \alpha \nu^2 \sigma_{\varphi_{it}}^2 \\
\mathrm{V}\,(\zeta_{it}) &= \alpha^2 \omega_t^2 \sigma_{\eta_{it}}^2 + \sigma_{u_{it}}^2 = \alpha^2 \mathrm{V}\,(\xi_{it}) + \sigma_{u_{it}}^2 - \alpha^2 \nu^2 \sigma_{\varphi_{it}}^2 \,.
\end{aligned}
$$

The asymptotic biases B_B and B_W on the between and within regressions (OLS) estimates of the α parameter are thus respectively:

$$B_B = \lambda \frac{\sigma^2_{\mu_i}}{V(\xi_i)} \qquad B_W = -\alpha\nu^2 \frac{\sigma^2_{\varphi_{it}}}{V(\xi_{it})}\,.$$

We also introduce x-related heteroscedasticity in the regression model by assuming that the slope coefficient α is itself randomly distributed across individuals. We thus assume:

$$\alpha_i = \alpha + \gamma_i\,,$$

where γ_i is a normal random variable with mean zero and standard deviation σ_{α_i}, independent from the other random variables generating x. The disturbance in the regression thus becomes: $\widetilde{v}_{it} = v_{it} + \gamma_i x_{it}$, and its covariance matrix conditional on x can be written as

$$E(\underline{\widetilde{v}}_i\underline{\widetilde{v}}_i'|x_{it}) = E(\underline{v}_i\underline{v}_i') + \sigma^2_{\gamma_i}.E\left(\underline{x}_i\underline{x}_i'\right)\,.$$

5.10.3 Calibration of Simulations

We calibrate the simulations so that $V(\xi_i) = \sigma^2_{\mu_i}$ and $V(\xi_{it}) = \sigma^2_{u_{it}} = \sigma^2_{\varphi_{it}}$. This also implies the following relations between the variances of the random variables η_i and the specification error parameters λ and ν:

$$\begin{aligned}\sigma^2_{\eta_i} &= (1-\lambda^2)\sigma^2_{\mu_i}\\ \sigma^2_{\eta_{it}} &= (1-\nu^2)\sigma^2_{u_{it}} = (1-\nu^2)\sigma^2_{\varphi_{it}}\end{aligned}$$

and the between and within biases become

$$B_B = \lambda \qquad B_W = -\alpha\nu^2.$$

We also normalize the total variance of x to be equal to 1 and impose the shares of the between and within variances to be respectively S_B and $S_W = (1-S_B)$, that is:

$$\begin{aligned}\mathrm{V}(x_{it}) &= \mathrm{V}(\xi_i) + \mathrm{V}(\xi_{it}) = 1\\ \mathrm{V}(\xi_i) &= S_B\\ \mathrm{V}(\xi_{it}) &= S_W = (1-S_B)\,.\end{aligned}$$

The precise value that we choose for the parameters are such as to reproduce some of the basic features found in real panel data sets used to estimate firm production functions as in Mairesse (1990), Mairesse (1990), Hall and Mairesse (1995 and 1996), and Griliches and Mairesse (1998). The true mean value of the parameter of interest is taken to be $\alpha = 0.5$ and its standard deviation over individuals to be $\sigma_{\alpha_i} = 0.2$; the shares of the between and within variances of x are respectively $S_B = 0.8$ and $S_W = 0.2$; the weights entering the definition of the true

within component ξ_{it}^* of x are obtained with $\rho = 0.7$ and $l = 5$, that is $w_0 = 0.719$, $w_1 = 0.503$, $w_2 = 0.352$, $w_3 = 0.247$, $w_4 = 0.173$ and $w_5 = 0.121$. The correlated effect parameter λ is chosen equal to 0.2 so that $B_B = 0.2$ and the errors in variables parameter is chosen equal to $\sqrt{0.2/0.5}$ so that $B_W = -0.2$.

5.10.4 Three Scenarios

Basically we consider the three scenarios of Non Correlated Errors (NCE), of Correlated Effects (CE) with a value of 0.2 for the between bias B_B, and of Correlated Effects and Errors in Variables (CE+EV), with a value of 0.2 for the between bias B_B and of -0.2 for the within bias B_W. We investigate these three scenarios for six combinations of cross–sectional and time series sample sizes: $N = 100$, 400, and 1600, and $T = 3$ and 6. For each of the eighteen (3×6) configurations of scenarios and sample sizes, we performed 100 replications. We also experimented with 1000 replications but found only small differences in the results.

5.10.5 The Chamberlain and GMM Estimators

For each of the three true specifications (NCE), (CE) and (CE+EV), we assume these three different specifications and also the Correlated Effects and Weak Simultaneity specification (CE+WS). For each of the twelve combinations of true and assumed specifications, we compute two Chamberlain estimators: CHAMB1 and CHAMB2 and two GMM estimators: GMM1 and GMM2.

The CHAMB1 and CHAMB2 estimators are based on estimating equations of the form $\pi^0 = H\left(m^0\right)\theta^0$. They differ in their weight matrices S_1 and S_2 which are consistent estimates of the inverse of the variance matrices of π, respectively assuming homoscedasticity and allowing for possible heteroscedasticity; i.e.,

$$\text{plim} S_1 = V_{\underline{\hat{\pi}}}^{c-1} = [E(\underline{w}_i \underline{w}_i') \otimes E(\underline{x}_i \underline{x}_i')^{-1}]^{-1};$$
$$\text{plim} S_2 = V_{\underline{\hat{\pi}}}^{-1} = [I \otimes E(\underline{x}_i \underline{x}_i')^{-1} E(\underline{w}_i \underline{w}_i' \otimes \underline{x}_i \underline{x}_i') I \otimes E(\underline{x}_i \underline{x}_i')^{-1}]^{-1}.$$

The matrix S_1 is non–optimal since we have introduced x related heteroscedasticity in the model. The matrix S_2 is optimal for the (NCE) and (CE) assumed specifications (in which $H\left(\underline{m}^0\right) = H$). It is not "fully" optimal for the (CE+EV) and (CE+WS) assumed specifcations, in which cases it is (in principle) necessary to take into account the estimation of $E\left(\underline{x}_i \underline{x}_i'\right)$. The fully optimal weight matrix is a consistent estimate of the inverse of the covariance matrix of estimating equations involving the gradient $D^0 = \frac{\partial H(\underline{m})\theta^0}{\partial \underline{m}}(\underline{m}^0)$.

The GMM1 and GMM2 estimators are based on orthogonality conditions that take the form $E(z_i' \underline{\varepsilon}_i) = 0$, where z_i is the appropriate matrix of instruments and $\underline{\varepsilon}_i$ stands for the residuals $(\underline{\varepsilon}_i = \underline{v}_i)$ when assuming (NCE), and for their first difference $(\underline{\varepsilon}_i = \Delta \underline{v}_i)$ when assuming (CE), (CE+EV) and (CE+WS). The weight

matrix used for GMM1 is $W_1 = \overline{z_i' z_i}^{-1}$ while the weight matrix used for GMM2 is $W_2 = \overline{z_i' \widehat{\varepsilon}_i \widehat{\varepsilon}_i' z_i}^{-1}$, where the $\widehat{\varepsilon}_i$ are the first step estimated residuals. As suggested by Arellano and Bond (1991), in the case of the (CE), (CE+EV) and (CE+WS) assumed specifications, we could also have used for the first step weight matrix an estimate of $E(z_i' D z_i)$ with $D = \Delta\Delta'$ (where Δ is the $(T-1)T$ difference matrix and D is the $(T-1)(T-1)$ matrix such that its diagonal coefficients are equal to 2 and its first upper and lower diagonal coefficients are equal to -1, and all other coefficients are zero). D is the optimal weight matrix in the cases of Correlated Effects if the time varying disturbances u_{it} are homoscedastic and serially uncorrelated.

5.10.6 Standard Errors and Specification Tests

The standard errors are derived from the covariance matrix of the estimated parameters which are computed using the following formulas:

$$\begin{array}{ll} \text{CHAMB1} & [H'S_1H]^{-1}H'S_1S_2^{-1}S_1H[H'S_1H]^{-1} \\ \text{CHAMB2} & [H'S_2H]^{-1} \\ \text{GMM1} & [G'W_1G]^{-1}G'W_1W_2^{-1}W_1[G'W_1G]^{-1} \\ \text{GMM2} & [G'W_2G]^{-1} \end{array}$$

where G is the gradient of the orthogonality conditions with respect to the parameter of interest, that is $G = E(z_i' \underline{x}_i)$ if (NCE), and $G = E(z_i' \Delta \underline{x}_i)$ if (CE), (CE+EV) and (CE+WS).

The specification tests are implemented using the CHAMB2 and GMM2 estimators. They are based on the following χ^2 statistics:

$$\begin{array}{ll} \text{CHAMB} & N\left(\widehat{\pi} - H(\widehat{m})\,\widehat{\theta}\right)' S_2 \left(\widehat{\pi} - H(\widehat{m})\,\widehat{\theta}\right) \\ \text{GMM} & N\left(\overline{z_i' \widehat{\underline{\varepsilon}}_i}\right)' W_2 \left(\overline{z_i' \widehat{\underline{\varepsilon}}_i}\right) \end{array}$$

Acknowledgments We are grateful to Lazlo Matyas and Patrick Sevestre for their energetic editorial help and friendly collaboration in the various stages of writing this chapter. We have also benefited from comments by G. Chamberlain, Z. Griliches, T. Klette, M. Lechner, and F. Windmeijer on its first version published in the second edition of this book, and by B. Hall, P. Mohnen and W. Raymond on the present revised version for this new edition.

References

Abowd J., B. Crepon, F. Kramarz and A. Trognon (1995): *A la Recherche des Moments Perdus, Covariance Models for Unbalanced Panels,* NBER Technical Working Paper n°180.

Andrews, D.W.K. (1985): *Asymptotic Results for Generalised Wald Tests,* Cowles Foundation D.P, 761R.

Ahn, S.C. and P. Schmidt (1995): Efficient Estimation of Models for Dynamic Panel Data, *Journal of Econometrics*, 68, 5–25.

Arellano, M. and S.R. Bond (1991): Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations, *Review of Economic Studies*, 58, 277–297.
Arellano, M. and O. Bover (1995): Another Look at the Instrumental Variable Estimation of Error-Components Models, *Journal of Econometrics*, 68, 29–52.
Berndt, E. and N. Savin (1977): Conflict among Criteria for Testing Hypotheses in the Multivariate Regression Model, *Econometrica*, 45, 1263–1278.
Bertscheck, I. and M. Lechner (1995): *GMM Estimation of Panel Probit Models: Nonparametric Estimation of the Optimal Instruments*, Humboldt Universitat zu Berlin, Discussion Paper N.25.
Blundell, R.W. and S. Bond (1998): Initial conditions and Moment Restrictions in Dynamic Panel Data Models, *Journal of Econometrics*, 87, 115–143.
Blundell, R.W. and S. Bond (2000): GMM Estimation with Highly Persistent Panel Data: An Application to Production Function Estimation, *Econometric Reviews*, 19, 321–340.
Bound, J, D.A. Jaeger, and R. Baker (1993): *The Cure Can Be Worse than the Disease: A Cautionary Tale Regarding Instrumental Variables*, National Bureau of Economic Research, Technical Working Paper N.137.
Breusch, T.S. (1979): Conflict Among Criteria for Testing Hypotheses: Extensions and Comments, *Econometrica*, 47, 203–207.
Chamberlain, G. (1982): Multivariate Regression Model for Panel Data, *Journal of Econometrics*, 18, 5–46.
Chamberlain, G. (1984): *Panel Data* in Handbook of Econometrics ed. by Z. Griliches and M.D. Intriligator, Amsterdam, North Holland Publishing Co, pp. 1247–1318.
Crépon, B., F. Kramarz and A. Trognon (1998): Parameters of Interest, Nuisance Parameters and Orthogonality Conditions: An Application to Autoregressive Error Component Models, *Journal of Econometrics*, 82, 135–156.
Gourieroux, C., A. Monfort and A. Trognon (1985): Moindres Carres Asymptotiques, *Annales de l'INSEE*, 58, 91–122.
Gourieroux, C. and A. Monfort (1989): A General Framework for Testing a Null Hypothesis in a Mixed Form, *Econometric Theory*, 5, 63–82.
Griliches, Z. and J. Hausman (1986): Errors in Variables in Panel Data, *Journal of Econometrics*, 31, 93–118.
Griliches, Z. and J. Mairesse (1998): *Production Functions: The Search for Identification*, in Econometrics and Economic Theory in the 20th Century : The Ragnar Frish Centennial Symposium ed. by S. Ström, Cambridge, Cambridge University Press, pp. 169–203.
Hall B.H.H and Mairesse, J. (1995): Exploring the Relationship between R-D and Productivity in French Manufacturing Firms, *Journal of Econometrics*, 65, 263–293.
Hall B.H.H and Mairesse, J. (1996): *Estimating the Productivity of Research and Development in French and United States Manufacturing Firms : An Exploration of Simultaneity Issues with GMM Methods*, in International Productivity Differences and Their Explanations ed. by K. Wagner and Bart Van Ark, Elsevier Science, pp. 285–315.
Holtz-Eakin, D., W.K. Newey, and H. Rosen (1988): Estimating Vector Autoregressions with Panel data, *Econometrica*, 56, 1371–1395.
Malinvaud, E. (1970): *Statistical Methods of Econometrics*, North Holland Publishing Co, Amsterdam.
Mundlak, Y. (1961): Empirical Production Function Free of Management Bias, *Journal of Farm Economics*, 43, 45–56.
Magnus, J.R. and H. Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley and Sons, Hoboken.
Mairesse, J. (1990): *Time series and Cross-sectional Estimates on Panel Data: Why are they Different and Why Should they Be Equal?* in Panel Data and Labor Market Studies ed. by J. Hartog et al., Amsterdam, North-Holland Publishing Co, pp. 81–95.
Mairesse, J. and Z. Griliches (1990): *Heterogeneity in Panel Data: Are There Stable Production Functions?* in Essays in Honour of Edmond Malinvaud ed. by P. Champsaurs et al., Cambridge, MIT Press, pp. 192–231.

Newey, W.K. (1985): Generalised Method of Moments Specification Testing, *Journal of Econometrics*, 29, 229–256.
Rubin, D.B. (1976): Inference and Missing Data, *Biometrika*, 63, 581–92.
Stewart G.W. (1973): *Introduction to Matrix Computations*, Academic Press, Orlando.
White, H. (1980): A Heteroskedasticity Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica* 48, 817–838.

Chapter 6
Random Coefficient Models

Cheng Hsiao and M. Hashem Pesaran

6.1 Introduction

Consider a linear regression model of the form

$$y = \boldsymbol{\beta}'\mathbf{x} + u, \tag{6.1}$$

where y is the dependent variable and $\mathbf{x}$ is a $K \times 1$ vector of explanatory variables. The variable u denotes the effects of all other variables that affect the outcome of y but are not explicitly included as independent variables. The standard assumption is that u behaves like a random variable and is uncorrelated with $\mathbf{x}$. However, the emphasis of panel data is often on the individual outcomes. In explaining human behavior, the list of relevant factors may be extended ad infinitum. The effect of these factors that have not been explicitly allowed for may be individual specific and time varying. In fact, one of the crucial issues in panel data analysis is how the differences in behavior across individuals and/or through time that are not captured by $\mathbf{x}$ should be modeled.

The variable intercept and/or error components models attribute the heterogeneity across individuals and/or through time to the effects of omitted variables that are individual time-invariant, like sex, ability and social economic background variables that stay constant for a given individual but vary across individuals, and/or period individual-invariant, like prices, interest rates and widespread optimism or pessimism that are the same for all cross-sectional units at a given point in time but vary through time. It does not allow the interaction of the individual specific and/or

Cheng Hsiao
University of Southern California and Nanyang Technological University, Singapore,
e-mail: chsiao@usc.edu

M. Hashem Pesaran
Cambridge University and USC, Sidgwick Avenue, Cambridge, CB3 9DD, United Kingdom,
e-mail: mhp1@econ.cam.ac.uk

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

time varying differences with the included explanatory variables, **x**. A more general formulation would be to let the variable y of the individual i at time t be denoted as

$$\begin{aligned} y_{it} &= \boldsymbol{\beta}_{it}' \mathbf{x}_{it} + u_{it}, \\ &= \beta_{1it} x_{1it} + \ldots + \beta_{kit} x_{kit} + u_{it}, \end{aligned} \tag{6.2}$$

$i = 1, \ldots, N$, and $t = 1, \ldots, T$. Expression (6.2) corresponds to the most general specification of the panel linear data regression problem. It simply states that each individual has their own coefficients that are specific to each time period. However, as pointed out by Balestra (1996) this general formulation is, at most, descriptive. It lacks any explanatory power and it is useless for prediction. Furthermore, it is not estimable as the number of parameters to be estimated exceeds the number of observations. For a model to become interesting and to acquire explanatory and predictive power, it is essential that some structure is imposed on its parameters.

One way to reduce the number of parameters in (6.2) is to adopt an analysis of variance framework by letting

$$\beta_{kit} = \beta_k + \alpha_{ki} + \lambda_{kt}, \sum_{i=1}^{N} \alpha_{ki} = 0, \text{ and } \sum_{t=1}^{T} \lambda_{kt} = 0, \; k = 1, \ldots, K. \tag{6.3}$$

This specification treats individual differences as fixed and is computationally simple. The drawback is that it is not parsimonious, and hence reliable estimates of α_{ki} and λ_{kt} are difficult to obtain. Moreover, it is difficult to draw inference about the population if differences across individuals and/or over time are fixed and different.

An alternative to the fixed coefficient (or effects) specification of (6.3) is to let α_{ki} and λ_{kt} be random variables and introduce proper stochastic specifications. This is commonly called the "random coefficients" model. The random coefficient specification reduces the number of parameters to be estimated substantially, while still allowing the coefficients to differ from unit to unit and/or from time to time.

In Sect. 6.2 we introduce various types of random coefficients models and suggest a common framework for them. In Sects. 6.3 and 6.4 we consider the fundamental issues of statistical inference of a random coefficients formulation using the sampling approach. In Sect. 6.5 we consider a Bayesian approach. Section 6.6 considers the generalization to a dynamic framework. Issues of testing for homogeneity under weak exogeneity are discussed in Sect. 6.7. Discussions on random coefficients, simultaneous equation systems and cross-sectional dependence are provided in Sects. 6.8 and 6.9. Conclusions are in Sect. 6.10.

6.2 The Models

Let there be observations for N cross-sectional units over T time periods. Suppose the variable y for the ith unit at time t is specified as a linear function of K strictly exogenous variables, $x_{kit}, k = 1, 2, \ldots, K$, in the form[1]

[1] The case where one or more of the regressors are weakly exogenous is considered in Sect. 6.6.

$$\begin{aligned} y_{it} &= \sum_{k=1}^{K} \beta_{kit} x_{kit} + u_{it}, \\ &= \boldsymbol{\beta}_{it}' \mathbf{x}_{it} + u_{it},\ i = 1 \ldots, N,\ t = 1, \ldots, T, \end{aligned} \tag{6.4}$$

where u_{it} denotes the random error term, $\mathbf{x}_{it}$ is a $K \times 1$ vector of exogenous variables and $\boldsymbol{\beta}_{it}$ is the $K \times 1$ vector of coefficients. The random coefficients approach assumes that the coefficients $\boldsymbol{\beta}_{it}$ are drawn from probability distributions with a fixed number of parameters that do not vary with N and/or T. Depending on the type of assumption about the parameter variation, we can further classify the models into one of two categories: stationary and non-stationary random-coefficients models.

The stationary random-coefficients models regard the coefficients as having constant means and variance-covariances. Namely, the $K \times 1$ vector $\boldsymbol{\beta}_{it}$ is specified as

$$\boldsymbol{\beta}_{it} = \boldsymbol{\beta} + \boldsymbol{\xi}_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{6.5}$$

where $\boldsymbol{\beta}$ is a $K \times 1$ vector of constants, and $\boldsymbol{\xi}_{it}$ is a $K \times 1$ vector of stationary random variables with zero means and constant variance–covariances. For instance, in the Swamy (1970) type random coefficient models,

$$\boldsymbol{\beta}_{it} = \boldsymbol{\beta} + \boldsymbol{\alpha}_i, \ \ i = 1, \ldots, N, \ \ t = 1, \ldots, T, \tag{6.6}$$

and

$$\begin{aligned} E(\boldsymbol{\alpha}_i) = \mathbf{0},\ E(\boldsymbol{\alpha}_i \mathbf{x}_{it}') = \mathbf{0}, \\ E(\boldsymbol{\alpha}_i \boldsymbol{\alpha}_j') = \begin{cases} \boldsymbol{\Delta}, \text{ if } i = j, \\ \mathbf{0}, \text{ if } i \neq j. \end{cases} \end{aligned} \tag{6.7}$$

Hsiao (1974, 1975) considers the following type of model

$$\begin{aligned} \boldsymbol{\beta}_{it} &= \boldsymbol{\beta} + \boldsymbol{\xi}_{it} \\ &= \boldsymbol{\beta} + \boldsymbol{\alpha}_i + \lambda_t, \quad i = 1, \ldots, N,\ t = 1, \ldots, T, \end{aligned} \tag{6.8}$$

and assumes

$$\begin{aligned} E(\boldsymbol{\alpha}_i) = E(\lambda_t) = \mathbf{0},\ E\left(\boldsymbol{\alpha}_i \lambda_t'\right) = \mathbf{0}, \\ E\left(\boldsymbol{\alpha}_i \mathbf{x}_{it}'\right) = \mathbf{0},\ E\left(\lambda_t \mathbf{x}_{it}'\right) = \mathbf{0}, \\ E\left(\boldsymbol{\alpha}_i \boldsymbol{\alpha}_j'\right) = \begin{cases} \boldsymbol{\Delta}, \text{ if } i = j, \\ \mathbf{0}, \text{ if } i \neq j, \end{cases} \\ E(\boldsymbol{\lambda}_t \boldsymbol{\lambda}_s') = \begin{cases} \boldsymbol{\Lambda}, \text{ if } t = s, \\ \mathbf{0}, \text{ if } t \neq s. \end{cases} \end{aligned} \tag{6.9}$$

Alternatively, a time varying parameter model may be treated as realizations of a stationary stochastic process, thus $\boldsymbol{\beta}_{it}$ can be written in the form,

$$\boldsymbol{\beta}_{it} = \boldsymbol{\beta}_t = \mathbf{H} \beta_{t-1} + \boldsymbol{\eta}_t, \tag{6.10}$$

where all eigenvalues of $\mathbf{H}$ lie inside the unit circle, and $\boldsymbol{\eta}_t$ is a stationary random variable with mean $\boldsymbol{\mu}$. Then the Hildreth and Houck (1968) type model is obtained by letting $\mathbf{H} = \mathbf{0}$ and $\boldsymbol{\eta}_t$ be i.i.d.; for the Pagan (1980) model, $\mathbf{H} = \mathbf{0}$ and

$$\boldsymbol{\eta}_t - \boldsymbol{\mu} = \boldsymbol{\eta}_t - \overline{\boldsymbol{\beta}} = a(L)\varepsilon_t, \tag{6.11}$$

where $\overline{\boldsymbol{\beta}}$ is the mean of $\boldsymbol{\beta}_t$ and $a(L)$ is the ratio of polynomials of orders p and q in the lag operator $L(L\varepsilon_t = \varepsilon_{t-1})$ and ε_t is independent normal. The Rosenberg (1972, 1973) return-to-normality model assumes the absolute value of the characteristic roots of $\mathbf{H}$ be less than 1 with $\boldsymbol{\eta}_t$ independently normally distributed with mean $\boldsymbol{\mu} = (\mathbf{I}_K - \mathbf{H})\overline{\boldsymbol{\beta}}$.

The nonstationary random coefficients models do not regard the coefficient vector as having constant mean or variances. Changes in coefficients from one observation to the next can be the result of the realization of a nonstationary stochastic process or can be a function of exogenous variables. When the coefficients are realizations of a nonstationary stochastic process, we may again use (6.10) to represent such a process. For instance, the Cooley and Prescott (1976) model can be obtained by letting $\mathbf{H} = \mathbf{I}_K$ and $\boldsymbol{\mu} = \mathbf{0}$. When the coefficients $\boldsymbol{\beta}_{it}$ are functions of individual characteristics or time variables (e.g. Amemiya (1978), Boskin and Lau (1990), Hendricks, Koenker, and Poirier (1979), Singh, Nagar, Choudhry and Raj (1976), Swamy and Tinsley (1977) and Wachter (1976)) we can let

$$\boldsymbol{\beta}_{it} = \boldsymbol{\Gamma}\mathbf{q}_{it} + \boldsymbol{\eta}_{it}. \tag{6.12}$$

While the detailed formulation and estimation of the random coefficients model depends on the specific assumptions about the parameter variation, many types of the random coefficients models can be conveniently represented using a mixed fixed and random coefficients framework of the form (e.g. Hsiao (1990) and Hsiao, Appelbe and Dineen (1992))

$$\mathbf{y}_{it} = \mathbf{z}_{it}'\bar{\gamma} + \mathbf{w}_{it}'\boldsymbol{\alpha}_{it} + u_{it},\ i = 1,\ldots,N,\ \ t = 1,\ldots,T, \tag{6.13}$$

where $\mathbf{z}_{it}$ and $\mathbf{w}_{it}$ are vectors of exogenous variables with dimensions ℓ and p respectively, $\bar{\gamma}$ is an $\ell \times 1$ vector of constants, $\boldsymbol{\alpha}_{it}$ is a $p \times 1$ vector of random variables, and u_{it} is the error term. For instance, the Swamy type model ((6.6) and (6.7)) can be obtained from (6.13) by letting $\mathbf{z}_{it} = \mathbf{w}_{it} = \mathbf{x}_{it}, \bar{\gamma} = \overline{\beta}$, and $\boldsymbol{\alpha}_{it} = \boldsymbol{\alpha}_i$; the Hsiao type model (6.8) and (6.9) is obtained by letting $\mathbf{z}_{it} = \mathbf{w}_{it} = \mathbf{x}_{it}, \bar{\gamma} = \overline{\boldsymbol{\beta}}$, and $\boldsymbol{\alpha}_{it} = \boldsymbol{\alpha}_i + \boldsymbol{\lambda}_t$; the stochastic time varying parameter model (6.10) is obtained by letting $\mathbf{z}_{it} = \mathbf{x}_{it}$, $\mathbf{w}_{it}' = \mathbf{x}_{it}'[\mathbf{H}, \mathbf{I}_K]$, $\bar{\gamma} = \boldsymbol{\mu}$, and $\boldsymbol{\alpha}_{it}' = \lambda_t' = [\boldsymbol{\beta}_{t-1}', (\boldsymbol{\eta}_t - \boldsymbol{\mu})']$; and the model where $\boldsymbol{\beta}_{it}$ is a function of other variables (6.12) is obtained by letting $\mathbf{z}_{it}' = \mathbf{x}_{it}' \otimes \mathbf{q}_{it}'$, $\bar{\gamma}' = \mathrm{vec}(\Gamma)$, $\mathbf{w}_{it} = \mathbf{x}_{it}$, $\alpha_{it} = \boldsymbol{\eta}_{it}$, etc.

For ease of illustrating the fundamental issues involved in estimating a random coefficients model we shall make the simplifying assumption that $\boldsymbol{\alpha}_{it} = \boldsymbol{\alpha}_i$ and $\boldsymbol{\alpha}_i$ are independently normally distributed over i with mean $\mathbf{0}$ and covariance $\boldsymbol{\Delta}$, denoted

by $\boldsymbol{\alpha}_i \sim N(\mathbf{0}, \Delta)$.[2] In other words, there are only individual-specific effects, $\boldsymbol{\alpha}_i$, and these individual-specific effects stay constant over time. Under this simplified assumption, model (6.13) can be written in the stacked form

$$\mathbf{y} = \overline{\mathbf{Z}}\bar{\gamma} + \mathbf{W}\alpha + \mathbf{u}, \tag{6.14}$$

where

$$\underset{NT\times 1}{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}, \ \underset{T\times 1}{\mathbf{y}_i} = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}, \ \underset{NT\times 1}{\mathbf{u}} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_N \end{pmatrix}, \ \underset{T\times 1}{\mathbf{u}_i} = \begin{pmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{pmatrix},$$

$$\underset{NT\times \ell}{\overline{\mathbf{Z}}} = \begin{pmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_N \end{pmatrix}, \ \underset{T\times \ell}{\mathbf{Z}_i} = \begin{pmatrix} \mathbf{z}'_{i1} \\ \vdots \\ \mathbf{z}'_{iT} \end{pmatrix},$$

$$\underset{NT\times Np}{\mathbf{W}} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 & \cdots & \mathbf{0} \\ \vdots & & \ddots & \\ \mathbf{0} & & & \mathbf{W}_N \end{pmatrix}, \ \underset{T\times p}{\mathbf{W}_i} = \begin{pmatrix} \mathbf{w}'_{i1} \\ \vdots \\ \mathbf{w}'_{iT} \end{pmatrix}, \text{ and } \underset{Np\times 1}{\boldsymbol{\alpha}} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_N \end{pmatrix}. \tag{6.15}$$

We further assume that $\boldsymbol{\alpha}$ and $\mathbf{u}$ are mutually independent with

$$E(\mathbf{u}) = \mathbf{0}, \text{ and } E\left(\mathbf{u}\mathbf{u}'\right) = \mathbf{C}. \tag{6.16}$$

6.3 Sampling Approach

Let

$$\mathbf{v} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{u}, \tag{6.17}$$

then $E(\mathbf{v}) = \mathbf{0}$ and

$$E\left(\mathbf{v}\mathbf{v}'\right) = \mathbf{W}(\mathbf{I}_N \otimes \boldsymbol{\Delta})\mathrm{W}' + \mathbf{C} = \Omega. \tag{6.18}$$

Model (6.14) can be viewed as a linear regression model of the form

$$\mathbf{y} = \overline{\mathbf{Z}}\bar{\gamma} + \mathbf{v}, \tag{6.19}$$

where the composite error term, $\mathbf{v}$, has a nonspherical covariance matrix. From a sampling point of view, the interest for model (6.19) will lie in (a) estimating the

[2] A model allowing the coefficients to vary across individuals and over time is very difficult to estimate. So far, most random coefficients models either assume $\boldsymbol{\alpha}_{it} = \boldsymbol{\alpha}_i$ or $\boldsymbol{\alpha}_{it} = \boldsymbol{\alpha}_t$. Here we shall only focus on the former. For the case of $\boldsymbol{\alpha}_{it} = \boldsymbol{\alpha}_t$, as in (6.10), one can employ Kalman filter type procedures to obtain MLE and carry out predictions. For details see Hsiao (2003).

mean coefficient vector $\bar{\gamma}$, (b) estimating the covariance matrix of $\mathbf{v}$, Ω, and (c) predicting y_{it}.

If $\mathbf{\Delta}$ and $\mathbf{C}$ are known, the best linear unbiased estimator of $\bar{\gamma}$ is the generalized least squares (GLS) estimator

$$\widehat{\bar{\gamma}} = (\overline{\mathbf{Z}}'\Omega^{-1}\overline{\mathbf{Z}})^{-1}(\overline{\mathbf{Z}}'\Omega^{-1}\mathbf{y}), \tag{6.20}$$

with covariance matrix

$$\mathbf{D} = \text{ Cov } (\widehat{\bar{\gamma}}) = (\overline{\mathbf{Z}}'\Omega^{-1}\overline{\mathbf{Z}})^{-1}. \tag{6.21}$$

If $\mathbf{\Delta}$ and $\mathbf{C}$ are unknown, we can apply a two step GLS estimator. In the first step we estimate $\mathbf{\Delta}$ and $\mathbf{C}$. In the second step we estimate $\bar{\gamma}$ by substituting the estimated $\mathbf{\Delta}$ and $\mathbf{C}$ into (6.20) and treating them as if they were known. Provided $\mathbf{\Delta}$ and $\mathbf{C}$ can be consistently estimated, the two step GLS will have the same asymptotic efficiency as the GLS estimator.

Similarly, we can obtain the best linear unbiased predictor of y_{if} using the formula

$$\begin{aligned} \widehat{y}_{if} &= \mathbf{z}'_{if}\bar{\gamma} + E(\upsilon_{if} \mid \mathbf{v}), \\ &= \mathbf{z}'_{if}\bar{\gamma} + \text{ Cov } (\upsilon_{if}, \mathbf{v})' \text{ Var } (\mathbf{v})^{-1}\mathbf{v}. \end{aligned} \tag{6.22}$$

Because $\bar{\gamma}$ and $\mathbf{v}$ are unknown, their estimated values, $\widehat{\bar{\gamma}}$ and $\hat{\mathbf{v}} = \mathbf{y} - \overline{\mathbf{Z}}\widehat{\bar{\gamma}}$ are substituted into (6.22) in practice.

Equations (6.20)–(6.22) provide a general principle for efficient inference of a random coefficients model. To illustrate relations to a specific type of random coefficients model, we consider a Swamy type model (6.4), (6.6) and (6.7), assuming that the regressors $\mathbf{z}_{it}$, are strictly exogenous.[3]

Under the assumptions of Swamy (1970), we have

$$\overline{\mathbf{Z}} = \mathbf{XA}, \mathbf{W} = \mathbf{X}, \ \bar{\gamma} = \overline{\boldsymbol{\beta}}, \ \boldsymbol{\beta}_i = \overline{\boldsymbol{\beta}} + \boldsymbol{\alpha}_i, \tag{6.23}$$

where

$$\underset{NT\times K}{\mathbf{A}} = (\mathbf{I}_K, \mathbf{I}_K, .., \mathbf{I}_K)' \tag{6.24}$$

$$\underset{NT\times K}{\mathbf{X}} = \begin{pmatrix} \mathbf{X}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{X}_N \end{pmatrix}, \ \underset{T\times K}{\mathbf{X}_i} = \begin{pmatrix} \mathbf{x}'_{i1} \\ \vdots \\ \mathbf{x}'_{iT} \end{pmatrix}.$$

For simplicity, we also assume that u_{it} is *independently* distributed across i and over t with

$$E\left(u_{it}^2\right) = \sigma_i^2. \tag{6.25}$$

[3] For estimation of correlated random coeffcient model using the instrumental variables approach, see Murtazashvili and Wooldridge (2007).

Then Ω is block diagonal, with the ith diagonal block equal to

$$\mathbf{\Omega}_i = \mathbf{X}_i \mathbf{\Delta} \mathbf{X}_i' + \sigma_i^2 \mathbf{I}_T. \tag{6.26}$$

Substituting (6.23)–(6.26) into (6.20), the best linear unbiased estimator of the mean coefficient vector $\overline{\beta}$ is

$$\begin{aligned} \hat{\overline{\beta}}_{\text{GLS}} &= \left(\mathbf{A}'\mathbf{X}'\Omega^{-1}\mathbf{X}\mathbf{A}\right)^{-1}\mathbf{A}'\mathbf{X}\Omega^{-1}\mathbf{y}, \\ &= \left(\sum_{i=1}^{N}\mathbf{X}_i'\Omega_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i'\Omega_i^{-1}\mathbf{y}_i\right), \\ &= \sum_{i=1}^{N}\mathbf{R}_i\hat{\beta}_i, \end{aligned} \tag{6.27}$$

where

$$\mathbf{R}_i = \left[\sum_{i=1}^{N}\left(\mathbf{\Delta} + \Sigma_{\hat{\beta}_i}\right)^{-1}\right]^{-1}\left(\mathbf{\Delta} + \Sigma_{\hat{\beta}_i}\right)^{-1}, \tag{6.28}$$

and

$$\hat{\beta}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{y}_i,\ \Sigma_{\hat{\beta}_i} = V\left(\hat{\beta}_i\right) = \sigma_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}. \tag{6.29}$$

The last expression of (6.27) is obtained by repeatedly utilizing the identity relation,

$$(\mathbf{E} + \mathbf{BFB}')^{-1} = \mathbf{E}^{-1} - \mathbf{E}^{-1}\mathbf{B}(\mathbf{B}'\mathbf{E}^{-1}\mathbf{B} + \mathbf{F}^{-1})^{-1}\mathbf{B}'\mathbf{E}^{-1}. \tag{6.30}$$

It shows that the GLS estimator is a matrix weighted average of the least squares estimator for each cross-sectional unit (6.29), with the weights inversely proportional to their covariance matrices. It also shows that the GLS estimator requires only a matrix inversion of order K, and so it is not much more complicated to compute than the sample least squares estimator.

The covariance matrix of the GLS estimator is

$$\text{Cov}\left(\hat{\overline{\beta}}_{\text{GLS}}\right) = \left(\mathbf{A}'\mathbf{X}'\Omega^{-1}\mathbf{X}\mathbf{A}\right)^{-1} = \left[\sum_{i=1}^{N}\mathbf{X}_i'\Omega_i^{-1}\mathbf{X}_i\right]^{-1} = \left[\sum_{i=1}^{N}\left(\Delta + \Sigma_{\hat{\beta}_i}\right)^{-1}\right]^{-1}. \tag{6.31}$$

If both errors and $\boldsymbol{\alpha}_i$ are normally distributed, the GLS estimator of $\overline{\boldsymbol{\beta}}$ is the maximum likelihood estimator (MLE) of $\overline{\boldsymbol{\beta}}$ conditional on $\mathbf{\Delta}$ and σ_i^2. Without knowledge of $\mathbf{\Delta}$ and σ_i^2, we can estimate $\overline{\boldsymbol{\beta}}$, $\mathbf{\Delta}$ and σ_i^2, $i = 1, \ldots, N$ simultaneously by the maximum likelihood method. However, computationally it can be tedious. A natural alternative is to first estimate Ω_i then substitute the estimated Ω_i into (6.27).

Swamy proposes using the least squares estimator of $\boldsymbol{\beta}_i, \hat{\boldsymbol{\beta}}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{y}_i$ and residuals $\widehat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\beta}_i$ to obtain unbiased estimators of $\sigma_i^2, i = 1, \ldots, N$, and $\mathbf{\Delta}$. Noting that

$$\widehat{\mathbf{u}}_i = [\mathbf{I}_T - \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i']\mathbf{u}_i, \tag{6.32}$$

and

$$\hat{\boldsymbol{\beta}}_i = \boldsymbol{\beta}_i + (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{u}_i, \tag{6.33}$$

we obtain the unbiased estimators of σ_i^2 and $\boldsymbol{\Delta}$ as:

$$\begin{aligned}\widehat{\sigma}_i^2 &= \frac{\widehat{\mathbf{u}}_i'\widehat{\mathbf{u}}_i}{T-K}, \\ &= \frac{1}{T-K}\mathbf{y}_i'[\mathbf{I}_T - \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i']\mathbf{y}_i,\end{aligned} \tag{6.34}$$

$$\begin{aligned}\widehat{\boldsymbol{\Delta}} = \frac{1}{N-1}\sum_{i=1}^{N}&\left(\hat{\beta}_i - N^{-1}\sum_{j=1}^{N}\hat{\beta}_j\right) \\ &\left(\hat{\beta}_i - N^{-1}\sum_{j=1}^{N}\hat{\beta}_j\right)' - \frac{1}{N}\sum_{i=1}^{N}\widehat{\sigma}_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}.\end{aligned} \tag{6.35}$$

Just as in the error-components model, the estimator (6.35) is not necessarily nonnegative definite. In this situation, Swamy [also see Judge, Griffiths, Hill, Lütkepohl and Lee (1985)] has suggested replacing (6.35) by

$$\widehat{\boldsymbol{\Delta}}^* = \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\beta}_i - N^{-1}\sum_{j=1}^{N}\hat{\beta}_j\right)\left(\hat{\beta}_i - N^{-1}\sum_{j=1}^{N}\hat{\beta}_j\right)'. \tag{6.36}$$

This estimator, although biased, is nonnegative definite and consistent when T tends to infinity.

6.4 Mean Group Estimation

A consistent estimator of $\overline{\boldsymbol{\beta}}$ can also be obtained under more general assumptions concerning $\boldsymbol{\beta}_i$ and the regressors. One such possible estimator is the Mean Group (MG) estimator proposed by Pesaran and Smith (1995) for estimation of dynamic random coefficient models. The MG estimator is defined as the simple average of the OLS estimators, $\hat{\boldsymbol{\beta}}_i$:

$$\widehat{\overline{\beta}}_{\mathrm{MG}} = N^{-1}\sum_{i=1}^{N}\hat{\beta}_i. \tag{6.37}$$

When the regressors are strictly exogenous and the errors, u_{it} are independently distributed, an unbiased estimator of the covariance matrix of $\widehat{\overline{\beta}}_{\mathrm{MG}}$ can be computed as

$$\widehat{\mathrm{Cov}}\left(\widehat{\overline{\beta}}_{\mathrm{MG}}\right) = N^{-1}\widehat{\boldsymbol{\Delta}}^*,$$

where $\widehat{\boldsymbol{\Delta}}^*$ is given by (6.36). For a proof first note that under the random coefficient model we have

$$\hat{\boldsymbol{\beta}}_i = \boldsymbol{\beta} + \boldsymbol{\alpha}_i + \boldsymbol{\xi}_i,$$

where

$$\boldsymbol{\xi}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{u}_i,$$

and

$$\hat{\bar{\beta}}_{\mathrm{MG}} = \boldsymbol{\beta} + \bar{\boldsymbol{\alpha}} + \bar{\boldsymbol{\xi}},$$

where $\overline{\boldsymbol{\alpha}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\alpha}_i$ and $\overline{\boldsymbol{\xi}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\xi}_i$. Therefore

$$\hat{\boldsymbol{\beta}}_i - \hat{\bar{\beta}}_{\mathrm{MG}} = (\boldsymbol{\alpha}_i - \overline{\boldsymbol{\alpha}}) + \left(\boldsymbol{\xi}_i - \overline{\boldsymbol{\xi}}\right),$$

$$\begin{aligned}\left(\hat{\boldsymbol{\beta}}_i - \hat{\bar{\beta}}_{\mathrm{MG}}\right)\left(\hat{\boldsymbol{\beta}}_i - \hat{\bar{\beta}}_{\mathrm{MG}}\right)' &= (\boldsymbol{\alpha}_i - \overline{\boldsymbol{\alpha}})(\boldsymbol{\alpha}_i - \overline{\boldsymbol{\alpha}})' + \left(\boldsymbol{\xi}_i - \overline{\boldsymbol{\xi}}\right)\left(\boldsymbol{\xi}_i - \overline{\boldsymbol{\xi}}\right)' \\ &\quad + (\boldsymbol{\alpha}_i - \overline{\boldsymbol{\alpha}})\left(\boldsymbol{\xi}_i - \overline{\boldsymbol{\xi}}\right)' + \left(\boldsymbol{\xi}_i - \overline{\boldsymbol{\xi}}\right)(\boldsymbol{\alpha}_i - \overline{\boldsymbol{\alpha}})',\end{aligned}$$

and

$$\sum_{i=1}^{N} E\left[\left(\hat{\boldsymbol{\beta}}_i - \hat{\bar{\beta}}_{\mathrm{MG}}\right)\left(\hat{\boldsymbol{\beta}}_i - \hat{\bar{\beta}}_{\mathrm{MG}}\right)'\right] = (N-1)\Delta + \left(1 - \frac{1}{N}\right)\sum_{i=1}^{N}\sigma_i^2\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}.$$

But

$$\begin{aligned}\mathrm{Cov}\left(\hat{\bar{\beta}}_{\mathrm{MG}}\right) &= \mathrm{Cov}\left(\overline{\boldsymbol{\alpha}}\right) + \mathrm{Cov}\left(\overline{\boldsymbol{\xi}}\right), \\ &= \frac{1}{N}\boldsymbol{\Delta} + \frac{1}{N^2}\sum_{i=1}^{N}\sigma_i^2 E\left[\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\right].\end{aligned}$$

Using the above results it is now easily seen that

$$E\left[\widehat{\mathrm{Cov}}\left(\hat{\bar{\beta}}_{\mathrm{MG}}\right)\right] = \mathrm{Cov}\left(\hat{\bar{\beta}}_{\mathrm{MG}}\right),$$

as required.

Finally, it is worth noting that the MG and the Swamy estimators are in fact algebraically equivalent for T sufficiently large.

6.5 Bayesian Approach

One can also derive the solutions for the model (6.14) from a Bayesian point of view. The Bayes approach assumes that all quantities, including the parameters, are random variables. Therefore, as part of the model, prior probability distributions are introduced for the parameters. The prior distribution is supposed to express a state of knowledge or ignorance about the parameters before the data is obtained.

The model (6.14) with the assumption that $\overline{\gamma}$ is fixed and $\boldsymbol{\alpha}_i$ is random, can be viewed as the state of knowledge about the parameters $\overline{\gamma}$ and $\boldsymbol{\alpha}$ before the data are obtained: The prior distributions of $\overline{\gamma}$ and α are independent. There is no information on $\overline{\gamma}$ but there is information on $\boldsymbol{\alpha}_i$, which is normally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Delta}$. This prior information is combined with the model (6.14) and data, $\mathbf{y}$ and $\mathbf{z}$, to revise the probability distribution of $\overline{\gamma}$ and $\boldsymbol{\alpha}$, which is called the posterior distribution. From this distribution inferences are made.

Formally, we assume that

A1. The prior distributions of $\overline{\gamma}$ and $\boldsymbol{\alpha}$ are independent, that is,

$$p(\overline{\gamma}, \boldsymbol{\alpha}) = p(\overline{\gamma}) \cdot p(\boldsymbol{\alpha}). \tag{6.38}$$

A2. There is no information about $\overline{\gamma}$,

$$p(\overline{\gamma}) \propto \text{ constant}. \tag{6.39}$$

A3. There is prior information about $\boldsymbol{\alpha}$,

$$\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{I}_N \otimes \Delta). \tag{6.40}$$

Theorem 6.1. *Suppose that, given $\overline{\gamma}$ and $\boldsymbol{\alpha}$,*

$$\mathbf{y} \sim N(\overline{\mathbf{Z}}\overline{\gamma} + \mathbf{W}\alpha, \mathbf{C}). \tag{6.41}$$

Under A1–A3,

(a) the marginal distribution of $\mathbf{y}$ *given* $\overline{\gamma}$ *is*

$$\mathbf{y} \sim N(\overline{\mathbf{Z}}\overline{\gamma}, \mathbf{C} + \mathbf{W}(\mathbf{I}_N \otimes \boldsymbol{\Delta})\mathbf{W}'), \tag{6.42}$$

(b) the distribution of $\overline{\gamma}$ *given* $\mathbf{y}$ *is* $N(\widehat{\overline{\gamma}}, \mathbf{D})$, *where* $\widehat{\overline{\gamma}}$ *and* $\mathbf{D}$ *are given by (6.20) and (6.21), respectively.*

(c) the distribution of $\boldsymbol{\alpha}$ *given* $\mathbf{y}$ *is* $N(\widehat{\boldsymbol{\alpha}}, \widetilde{\mathbf{D}})$, *where*

$$\begin{aligned} \widehat{\boldsymbol{\alpha}} = {} & \{\mathbf{W}'[\mathbf{C}^{-1} - \mathbf{C}^{-1}\overline{\mathbf{Z}}(\overline{\mathbf{Z}}'\mathbf{C}^{-1}\overline{\mathbf{Z}})^{-1}\overline{\mathbf{Z}}'\mathbf{C}^{-1}]\mathbf{W} + (\mathbf{I}_N \otimes \boldsymbol{\Delta}^{-1})\}^{-1} \\ & \cdot\{\mathbf{W}'[\mathbf{C}^{-1} - \mathbf{C}^{-1}\overline{\mathbf{Z}}(\overline{\mathbf{Z}}'\mathbf{C}^{-1}\overline{\mathbf{Z}})^{-1}\overline{\mathbf{Z}}'\mathbf{C}^{-1}]\mathbf{y}\}, \end{aligned} \tag{6.43}$$

and

$$\widetilde{\mathbf{D}} = \{\mathbf{W}'[\mathbf{C}^{-1} - \mathbf{C}^{-1}\overline{\mathbf{Z}}(\overline{\mathbf{Z}}'\mathbf{C}^{-1}\overline{\mathbf{Z}})^{-1}\overline{\mathbf{Z}}'\mathbf{C}^{-1}]\mathbf{W} + (\mathbf{I}_N \otimes \boldsymbol{\Delta}^{-1})\}^{-1}. \tag{6.44}$$

See Appendix A for a proof.

Recall that

$$\boldsymbol{\beta} = \mathbf{A}\overline{\boldsymbol{\beta}} + \boldsymbol{\alpha}, \tag{6.45}$$

and therefore the Bayes estimator of $\boldsymbol{\beta}$ can be obtained by substituting the Bayes estimators of $\overline{\boldsymbol{\beta}}$ and $\boldsymbol{\alpha}$ (6.27) and (6.43) into (6.45), namely:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{*} &= \mathbf{A}\widehat{\bar{\boldsymbol{\beta}}}_{\mathrm{GLS}} + \widehat{\boldsymbol{\alpha}} \\ &= \left(\mathbf{X}'\mathbf{C}^{-1}\mathbf{X} + \widetilde{\boldsymbol{\Delta}}^{-1}\right)^{-1}\left(\mathbf{X}'\mathbf{C}^{-1}\mathbf{y} + \widetilde{\boldsymbol{\Delta}}^{-1}\mathbf{A}\widehat{\bar{\boldsymbol{\beta}}}_{\mathrm{GLS}}\right),\end{aligned} \tag{6.46}$$

where

$$\widetilde{\boldsymbol{\Delta}} = \mathbf{I}_N \otimes \boldsymbol{\Delta}.$$

When $E\left(\mathbf{u}_i\mathbf{u}_j'\right) = \sigma_i^2\mathbf{I}_T$ if $i = j$, and $\mathbf{0}$ otherwise, as assumed by Swamy (1970), we have

$$\widehat{\boldsymbol{\beta}}_i^{*} = \left[\Sigma_{\hat{\beta}_i}^{-1} + \boldsymbol{\Delta}^{-1}\right]^{-1}\left[\Sigma_{\hat{\beta}_i}^{-1}\hat{\beta}_i + \boldsymbol{\Delta}^{-1}\widehat{\bar{\beta}}_{\mathrm{GLS}}\right], i = 1,2,\ldots,N \tag{6.47}$$

where

$$\hat{\beta}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{y}_i, \text{ and } \Sigma_{\hat{\beta}_i} = V(\hat{\beta}_i) = \sigma_i^2(\mathbf{X}_i'\mathbf{X}_i)^{-1}. \tag{6.48}$$

The Bayes estimator (6.47), is identical to the Lindley and Smith (1972) estimator for a linear hierarchical model. This is to be expected since the Swamy type assumptions and the Lindley–Smith linear hierarchical model are formally equivalent.

The above estimator can also be written as

$$\widehat{\boldsymbol{\beta}}_i^{*} = \mathbf{H}_i\hat{\beta}_i + (\mathbf{I}_K - \mathbf{H}_i)\widehat{\bar{\beta}}_{\mathrm{GLS}},$$

where

$$\mathbf{H}_i = \left[\Sigma_{\hat{\beta}_i}^{-1} + \boldsymbol{\Delta}^{-1}\right]^{-1}\Sigma_{\hat{\beta}_i}^{-1} = \boldsymbol{\Delta}\left(\boldsymbol{\Delta} + \Sigma_{\hat{\beta}_i}\right)^{-1}.$$

which shows that $\widehat{\boldsymbol{\beta}}_i^{*}$ is a weighted average of the OLS estimator, $\hat{\beta}_i$, and the Swamy estimator of $\bar{\beta}$. Also, $\mathbf{R}_i$ defined by (6.28) can be written as

$$\mathbf{R}_i = \left(\sum_{j=1}^{N}\mathbf{H}_j\right)^{-1}\mathbf{H}_i,$$

and hence

$$N^{-1}\sum_{i=1}^{N}\widehat{\boldsymbol{\beta}}_i^{*} = \sum_{i=1}^{N}\mathbf{R}_i\hat{\beta}_i = \widehat{\bar{\beta}}_{\mathrm{GLS}},$$

namely the simple mean of the Bayes estimators (which could be viewed as the Bayes Mean Group estimator) is equal to the Swamy estimator of $\bar{\boldsymbol{\beta}}$.

Remark 6.1. It is useful to put the random coefficients model in a Bayesian framework because many of the estimators based on the sampling approach can also be derived from the Bayes approach. For instance, as one can see from theorem 6.1(*b*) conditional on $\boldsymbol{\Delta}$ and $\mathbf{C}$, the Bayes estimator of $\overline{\gamma}$ for the model (6.14) is identical to the GLS estimator of $\overline{\gamma}$ (6.20). Furthermore, a Bayesian framework makes it clear

the role of prior knowledge or ignorance about the parameter $\boldsymbol{\theta} = (\overline{\gamma}, \boldsymbol{\alpha})$ given $\mathbf{y}$. The parameters $\boldsymbol{\theta}$ are treated as random variables and all probability statements are conditional. Ignorance about $\boldsymbol{\theta}$ would necessitate a specification of a diffuse prior to $\boldsymbol{\theta}$, which is typically specified as

$$p(\boldsymbol{\theta}) \propto \text{ constant.}$$

On the other hand, information about $\boldsymbol{\theta}$ would necessitate a specification of an informative prior. The Swamy type random coefficients formulation of $\boldsymbol{\beta}_i$ having mean $\overline{\boldsymbol{\beta}}$ and covariance $\boldsymbol{\Delta}$ is equivalent to specifying an informative prior for the parameters $\boldsymbol{\beta}_i$.

Remark 6.2. Typically, we use the expected value of an i.i.d. random variable as a predictor of the random variable. In panel data, we have two dimensions, a cross-sectional dimension and a time series dimension. Even though $\boldsymbol{\alpha}_i$ is assumed independently distributed across i, once a particular $\boldsymbol{\alpha}_i$ is drawn, it stays constant over time for the ith cross-sectional unit. Therefore, it makes sense to predict $\boldsymbol{\alpha}_i$, (for an example, see Hsiao, Mountain, Tsui and Luke Chan (1989)). The Bayes predictor of $\boldsymbol{\alpha}_i$ is different from the classical sampling approach predictor. For instance, for the Swamy type model the sampling approach predictor of $\boldsymbol{\beta}_i = \overline{\boldsymbol{\beta}} + \boldsymbol{\alpha}_i$ defined by (6.23) is the least squares estimator (6.48). The Bayes predictor of $\boldsymbol{\beta}_i$, given by (6.46) or (6.47), is a weighted average between the least squares estimator of $\boldsymbol{\beta}_i$ and the overall mean $\overline{\boldsymbol{\beta}}$. In other words, the Bayes estimator of the individual coefficients $\boldsymbol{\beta}_i$ "shrinks" the estimate of $\boldsymbol{\beta}_i$ based on the information of the ith individual (6.48) towards the grand mean $\overline{\boldsymbol{\beta}}$. An intuitive reason for doing so is because if the actual differences in $\boldsymbol{\beta}_i$ can be attributable to the work of chance mechanisms as postulated by de Finetti's (1964) exchangeability assumption, information about $\boldsymbol{\beta}_i$ can be obtained by examining the behaviour of others in addition to those of the ith cross-sectional unit because the expected value of $\boldsymbol{\beta}_i$ is the same as $\boldsymbol{\beta}_j$. When there are not many observations (i.e. T is small) with regard to the ith individual, information about $\boldsymbol{\beta}_i$ can be expanded by considering the responses of others. When T becomes large, more information about $\boldsymbol{\beta}_i$ becomes available and the weight gradually shifts towards the estimate based on the ith unit. As $T \to \infty$, the Bayes estimator approaches the least squares estimator $\hat{\boldsymbol{\beta}}_i$.

Remark 6.3. The derivation of the posterior distribution and the Bayes estimators $\overline{\gamma}$ and $\boldsymbol{\alpha}$ of model (6.14) is based on known $\mathbf{C}$ and $\boldsymbol{\Delta}$. When $\mathbf{C}$ and $\boldsymbol{\Delta}$ are unknown, in principle, we can first assign a joint prior of $\boldsymbol{\beta}, \overline{\boldsymbol{\beta}}, \mathbf{C}$ and $\boldsymbol{\Delta}$, and combine this with the likelihood function to obtain the joint posterior distribution. This distribution then has to be integrated with respect of $\mathbf{C}$ and $\boldsymbol{\Delta}$. In practice, this is most complex to execute. Lindley and Smith (1972), therefore, suggest to approximate the posterior distribution of $\overline{\gamma}$ and $\boldsymbol{\alpha}$ conditional on the modal value of $\boldsymbol{\Delta}$ and $\mathbf{C}$. The modal estimates of $\boldsymbol{\Delta}$ and $\mathbf{C}$ may be found by supposing $\overline{\gamma}$ and $\boldsymbol{\alpha}$ known, and then replacing $\overline{\gamma}$ and $\boldsymbol{\alpha}$ in the results by their modes. The sequence of iterations typically starts with assumed values of $\boldsymbol{\Delta}$ and $\mathbf{C}$ to calculate the mode of $\overline{\gamma}$ and $\boldsymbol{\alpha}$, say $\widehat{\overline{\gamma}}^{(1)}$ and $\widehat{\boldsymbol{\alpha}}^{(1)}$.

Treating $\widehat{\overline{\gamma}}^{(1)}$ and $\widehat{\boldsymbol{\alpha}}^{(1)}$ as known, we can find the mode for $\boldsymbol{\Delta}$ and $\mathbf{C}$, say $\widehat{\boldsymbol{\Delta}}^{(1)}$ and $\mathbf{C}^{(1)}$. The $\widehat{\boldsymbol{\Delta}}^{(1)}$ and $\mathbf{C}^{(1)}$ are then used to find $\overline{\gamma}^{(2)}$ and $\widehat{\boldsymbol{\alpha}}^{(2)}$, and so on.

For the Swamy type model (6.6) and (6.7) under the assumption that $\boldsymbol{\Delta}^{-1}$ has a Wishart distribution with p degrees of freedom and matrix $\mathbf{R}^*$, it is shown by Lindley and Smith (1972) that the mode estimator of $\boldsymbol{\Delta}$ is

$$\widehat{\boldsymbol{\Delta}} = \left\{ \mathbf{R}^* + \sum_{i=1}^{N} (\widehat{\boldsymbol{\beta}}_i^* - \widehat{\overline{\boldsymbol{\beta}}})(\widehat{\boldsymbol{\beta}}_i^* - \widehat{\overline{\boldsymbol{\beta}}})' \right\} / (N + p - K - 2). \tag{6.49}$$

6.6 Dynamic Random Coefficients Models

Because of the inertia in human behaviour or institutional or technological rigidity, often a behavioural equation is specified with lagged dependent variable(s) appearing as regressor(s). We will consider a dynamic model of the form

$$y_{it} = \rho_i y_{i,t-1} + \mathbf{x}_{it}' \boldsymbol{\beta}_i + u_{it}, \;\; i = 1, 2, \ldots, N; \;\; t = 1, 2, \ldots, T, \tag{6.50}$$

where $\mathbf{x}_{it}$ is a $K \times 1$ vector of exogenous variables, and the error term u_{it} is assumed to be independently, identically distributed over t with mean zero and variance σ_i^2, and is independent across i. Let $\boldsymbol{\theta}_i = (\rho_i, \boldsymbol{\beta}_i')'$. We assume that $\boldsymbol{\theta}_i$ is independently distributed across i with

$$E(\boldsymbol{\theta}_i) = \overline{\boldsymbol{\theta}} = \left(\overline{\rho}, \overline{\boldsymbol{\beta}}' \right)', \tag{6.51}$$

$$E\left[(\boldsymbol{\theta}_i - \overline{\boldsymbol{\theta}})(\boldsymbol{\theta}_i - \overline{\boldsymbol{\theta}})' \right] = \boldsymbol{\Delta}. \tag{6.52}$$

Rewrite $\boldsymbol{\theta}_i = \overline{\boldsymbol{\theta}} + \boldsymbol{\alpha}_i$, (6.51) and (6.52) are equivalent to

$$E(\boldsymbol{\alpha}_i) = \mathbf{0}, \;\; E\left(\boldsymbol{\alpha}_i \boldsymbol{\alpha}_j' \right) = \begin{cases} \boldsymbol{\Delta} \; if \; i = j, \\ \mathbf{0} \; if \; i \neq j. \end{cases} \tag{6.53}$$

Although we may maintain the assumption (6.9) that $E(\boldsymbol{\alpha}_i \mathbf{x}_{it}') = \mathbf{0}$, we can no longer assume that $E(\boldsymbol{\alpha}_i y_{i,t-1}) = \mathbf{0}$. Through continuous substitutions, we have

$$\begin{aligned} y_{i,t-1} = & \sum_{j=o}^{\infty} (\overline{\rho} + \alpha_{i1})^j \mathbf{x}_{i,t-j-1}' (\overline{\boldsymbol{\beta}} + \boldsymbol{\alpha}_{i2}) \\ & + \sum_{j=o}^{\infty} (\overline{\rho} + \alpha_{i1})^j u_{i,t-j-1}. \end{aligned} \tag{6.54}$$

It follows that $E(\boldsymbol{\alpha}_i y_{i,t-1}) \neq \mathbf{0}$.

The violation of the independence between the regressors and the individual effects $\boldsymbol{\alpha}_i$ implies that the pooled least squares regression of y_{it} on $y_{i,t-1}$, and $\mathbf{x}_{it}$ will yield inconsistent estimates of $\overline{\boldsymbol{\theta}}$, even for T and N sufficiently large. Pesaran and

Smith (1995) have noted that as $T \to \infty$, the least squares regression of y_{it} on $y_{i,t-1}$ and $\mathbf{x}_{it}$ yields a consistent estimator of $\boldsymbol{\theta}_i, \widehat{\boldsymbol{\theta}}_i$. They suggest a mean group estimator of $\overline{\boldsymbol{\theta}}$ by taking the average of $\widehat{\boldsymbol{\theta}}_i$ across i,

$$\widehat{\overline{\boldsymbol{\theta}}}_{\mathrm{MG}} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}_i. \tag{6.55}$$

The mean group estimator is consistent when both N and $T \to \infty$. In finite T, $\widehat{\boldsymbol{\theta}}_i$ for $\boldsymbol{\theta}_i$ is biased to the order of $1/T$. (Hurwicz (1950), Kiviet and Phillips (1993)) and the limited Monte Carlo appears to show that the mean group estimator can be severely biased when T is very small (Hsiao, Pesaran and Tahmiscioglu 1999). However, under the assumption that y_{i0} are fixed and known and $\boldsymbol{\alpha}_i$ and u_{it} are independently normally distributed, as discussed in Sect. 6.5 we can implement the Bayes estimator of $\widehat{\boldsymbol{\theta}}_i$ conditional on σ_i^2 and $\boldsymbol{\Delta}$,

$$\widehat{\overline{\boldsymbol{\theta}}}_B = \left\{\sum_{i=1}^{N}\left[\sigma_i^2(\mathbf{W}_i'\mathbf{W}_i)^{-1}+\boldsymbol{\Delta}\right]^{-1}\right\}^{-1}\sum_{i=1}^{N}\left[\sigma_i^2(\mathbf{W}_i'\mathbf{W}_i)^{-1}+\boldsymbol{\Delta}\right]\widehat{\theta}_i, \tag{6.56}$$

where here $\mathbf{W}_i = (\mathbf{y}_{i,-1}, \mathbf{X}_i)$ with $\mathbf{y}_{i,-1} = (y_{i0}, y_{i1}, \ldots, y_{iT-1})'$. This Bayes estimator is a weighted average of the least squares estimator of individual units with the weights being inversely proportional to individual variances. When $T \to \infty$, $N \to \infty$, and $N/T^{3/2} \to 0$, the Bayes estimator is asymptotically equivalent to the mean group estimator (6.55) (Hsiao et al. 1999).

In practice, the variance components, σ_i^2 and $\boldsymbol{\Delta}$ are rarely known. The Monte Carlo studies conducted by Hsiao et al. (1999) show that by following the approach of Lindley and Smith (1972) in assuming that the prior-distributions of σ_i^2 and $\boldsymbol{\Delta}$ are independent and are distributed as

$$P(\boldsymbol{\Delta}^{-1}, \sigma_1^2, \ldots, \sigma_n^2) = W(\boldsymbol{\Delta}^{-1}|(rR)^{-1}, r)\prod_{i=1}^{N}\sigma_i^{-2}, \tag{6.57}$$

yields a Bayes estimator almost as good as the Bayes estimator with known $\boldsymbol{\Delta}$ and σ_i^2, where $W(.)$ represents the Wishart distribution with scale matrix, rR, and degrees of freedom r (e.g. Anderson (1984)).

The Hsiao et al. (1999) Bayes estimator is derived under the assumption that the initial observation y_{i0} are fixed constants. As discussed in Anderson and Hsiao (1981, 1982), this assumption is clearly unjustifiable for a panel with finite T. However, contrary to the sampling approach where the correct modelling of initial observations is quite important, the Hsiao et al. (1999) Bayesian approach appears to perform fairly well in the estimation of the mean coefficients for dynamic random coefficient models as demonstrated in their Monte Carlo studies.

Remark 6.4. Model (6.50) has not imposed any constraint on the coefficient of the lag dependent variable, ρ_i. Often an investigator would like to impose the stability condition $|\rho_i| < 1$. One way to impose the stability condition on individual units

would be to assume that ρ_i follows a Beta distribution on (0,1). For a Bayes estimator under this assumption see Liu and Tiao (1980).

6.7 Testing for Heterogeneity Under Weak Exogeneity

Given the importance of heterogeneity, it is very important to test for it. There are at least three different categories of tests available: (i) direct tests of parameter equality of the type used by Zellner (1962) in a SURE framework; (ii) Hausman (1978) type tests of the difference between two estimators of θ (or its subset); or (iii) Swamy (1970) type tests based on the dispersion of individual slope estimates from a suitable pooled estimator. The first type of test is generally applicable when N is relatively small and T sufficiently large. Here we shall examine types (ii) and (iii), and assume that N and T are sufficiently large.

The Hausman method can be used in cases where it the two estimators are consistent under the null of homogeneity, whilst only one of them is efficient. Also, under the alternative hypothesis the two estimators converge to different values.

Denote the efficient estimator by subscript "e" and the inefficient but consistent estimator (under the alternative hypothesis) by the subscript "c". Then we have

$$V(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_e) = V(\widehat{\boldsymbol{\theta}}_c) - V(\widehat{\boldsymbol{\theta}}_e). \tag{6.58}$$

This is the result used by Hausman (1978) where it is assumed that $\widehat{\boldsymbol{\theta}}_e$ is asymptotically the *most* efficient estimator. However, it is easily shown that (6.58) hold under a weaker requirement, namely when the (asymptotic) efficiency of $\widehat{\boldsymbol{\theta}}_e$ cannot be enhanced by the information contained in $\widehat{\boldsymbol{\theta}}_c$. Consider a third estimator $\widehat{\boldsymbol{\theta}}_*$, defined as a convex combination of $\widehat{\boldsymbol{\theta}}_c$ and $\widehat{\boldsymbol{\theta}}_e$

$$\mathbf{q}'\widehat{\boldsymbol{\theta}}_* = (1-\delta)\mathbf{q}'\widehat{\boldsymbol{\theta}}_e + \delta\mathbf{q}'\widehat{\boldsymbol{\theta}}_c, \tag{6.59}$$

where $\mathbf{q}$ is a vector of constants, and δ is a scalar in the range $0 \leq \delta \leq 1$. Since, by assumption, the asymptotic efficiency of $\widehat{\boldsymbol{\theta}}_e$ cannot be enhanced by the knowledge of $\widehat{\boldsymbol{\theta}}_c$, then it must be that $V(\mathbf{q}'\widehat{\boldsymbol{\theta}}_*) \geq V(\mathbf{q}'\widehat{\boldsymbol{\theta}}_e)$, and hence the value of δ that minimises $V(\mathbf{q}'\widehat{\boldsymbol{\theta}}_*)$, say δ^*, should be zero. However, using (6.59) directly, we have

$$\delta^* = \frac{\mathbf{q}'[V(\widehat{\boldsymbol{\theta}}_e) - \mathrm{Cov}(\widehat{\boldsymbol{\theta}}_e, \widehat{\boldsymbol{\theta}}_c)]\mathbf{q}}{\mathbf{q}'V(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_e)\mathbf{q}} = 0, \tag{6.60}$$

and hence $\mathbf{q}'[V(\widehat{\boldsymbol{\theta}}_e) - \mathrm{Cov}(\widehat{\boldsymbol{\theta}}_e, \widehat{\boldsymbol{\theta}}_c)]\mathbf{q} = 0$. But, if this result is to hold for an arbitrary vector $\mathbf{q}$, we must have

$$V(\widehat{\boldsymbol{\theta}}_e) = \mathrm{Cov}(\widehat{\boldsymbol{\theta}}_e, \widehat{\boldsymbol{\theta}}_c). \tag{6.61}$$

Using this in

$$V(\widehat{\boldsymbol{\theta}}_c - \widehat{\boldsymbol{\theta}}_e) = V(\widehat{\boldsymbol{\theta}}_c) + V(\widehat{\boldsymbol{\theta}}_e) - 2\,\mathrm{Cov}(\widehat{\boldsymbol{\theta}}_e, \widehat{\boldsymbol{\theta}}_c),$$

yields (6.58) as desired.

In the context of testing for slope heterogeneity a number of different Hausman type tests can be used.

One possibility would be to compare the pooled estimator of $\overline{\boldsymbol{\theta}}$, defined by[4]

$$\widehat{\bar{\theta}}_{\mathrm{OLS}} = \left(\sum_{i=1}^{N} \mathbf{W}_i'\mathbf{W}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{W}_i'\mathbf{y}_i$$

with the mean group estimator $\widehat{\bar{\boldsymbol{\theta}}}_{\mathrm{MG}}$, defined by (6.55). When the focus of attention is on the mean long run coefficients $\boldsymbol{\delta}^* = E(\boldsymbol{\beta}_i/(1-\rho_i))$, as in Pesaran, Shin and Smith (1999) the heterogeneity test could be based directly on $\widehat{\boldsymbol{\delta}}^*_{\mathrm{OLS}} = \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}/(1-\hat{\rho}_{\mathrm{OLS}})$ and $\widehat{\boldsymbol{\delta}}^*_{\mathrm{MG}} = N^{-1}\sum_{i=1}^{N}\widehat{\boldsymbol{\delta}}^*_i$, where $\widehat{\boldsymbol{\delta}}^*_i = \hat{\beta}_i/(1-\hat{\rho}_i)$. Under the null of homogeneity the pooled and the mean group estimators are both consistent, although only the mean group estimator is consistent under the alternative hypothesis when lagged values of the dependent variables are included in the model.

Under the full homogeneity assumption ($\boldsymbol{\theta}_i = \boldsymbol{\theta}$, $\sigma_i^2 = \sigma^2$), the asymptotic variance matrices of the pooled and the mean group estimators (for a fixed N and a large T) are given by

$$\mathrm{Cov}(\sqrt{T}\widehat{\bar{\boldsymbol{\theta}}}_{\mathrm{OLS}}) = \frac{\sigma^2}{N}\left(N^{-1}\sum_{i=1}^{N}\boldsymbol{\Psi}_i\right)^{-1}, \tag{6.62}$$

and

$$\mathrm{Cov}(\sqrt{T}\widehat{\bar{\boldsymbol{\theta}}}_{\mathrm{MG}}) = \frac{\sigma^2}{N}\left(N^{-1}\sum_{i=1}^{N}\boldsymbol{\Psi}_i^{-1}\right), \tag{6.63}$$

where $\boldsymbol{\Psi}_i = p\lim_{T\to\infty}(\mathbf{W}_i'\mathbf{W}_i/T)$. Also we have

$$\mathrm{Cov}(\sqrt{T}\widehat{\bar{\boldsymbol{\theta}}}_{\mathrm{OLS}}, \sqrt{T}\widehat{\bar{\boldsymbol{\theta}}}_{\mathrm{MG}}) = \mathrm{Cov}(\sqrt{T}\widehat{\bar{\boldsymbol{\theta}}}_{\mathrm{OLS}})$$

thus directly establishing that

$$\mathrm{Cov}\left[\sqrt{T}\left(\widehat{\bar{\boldsymbol{\theta}}}_{\mathrm{MG}} - \widehat{\bar{\boldsymbol{\theta}}}_{\mathrm{OLS}}\right)\right] = \frac{\sigma^2}{N}\left\{\left(N^{-1}\sum_{i=1}^{N}\boldsymbol{\Psi}_i^{-1}\right) - \left(N^{-1}\sum_{i=1}^{N}\boldsymbol{\Psi}_i\right)^{-1}\right\},$$

which is a positive definite matrix, assuming that $\boldsymbol{\Psi}_i \neq \boldsymbol{\Psi}_j$, for some i and j.[5] This condition is generally satisfied when the model contains regressors with heterogeneous variances. The above results suggest the following statistic for testing the homogeneity hypothesis:

$$h = NT\left(\widehat{\bar{\boldsymbol{\theta}}}_{\mathrm{MG}} - \widehat{\bar{\boldsymbol{\theta}}}_{\mathrm{OLS}}\right)'\widehat{\mathbf{V}}^{-1}\left(\widehat{\bar{\boldsymbol{\theta}}}_{\mathrm{MG}} - \widehat{\bar{\boldsymbol{\theta}}}_{\mathrm{OLS}}\right),$$

[4] Similar exercises can also be carried out using fixed or random effects estimators. But to keep the exposition simple here we focus on pooled estimators.

[5] For a proof see the Appendix in Pesaran, Smith and Im (1996).

where

$$\widehat{\mathbf{V}} = \widehat{\sigma}^2_{\text{MG}} \left\{ N^{-1} \sum_{i=1}^{N} \left(\frac{\mathbf{W}_i' \mathbf{W}_i}{T} \right)^{-1} - \left(N^{-1} \sum_{i=1}^{N} \frac{\mathbf{W}_i' \mathbf{W}_i}{T} \right)^{-1} \right\}. \tag{6.64}$$

and $\widehat{\sigma}^2_{\text{MG}} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\sigma}^2_i$. In computing h, one could also equally use $\widehat{\sigma}^2_{\text{OLS}}$ instead of $\widehat{\sigma}^2_{\text{MG}}$. Under the null hypothesis

$$H_0 : \rho_i = \rho, \boldsymbol{\beta}_i = \boldsymbol{\beta}, \text{ and } \sigma_i^2 = \sigma^2, \text{ for all } i,$$

and for N and T sufficiently large we have

$$h \overset{a}{\sim} \chi^2_{K+1}.$$

When the focus of the analysis is on the long run coefficients we first note that[6]

$$\widehat{\boldsymbol{\delta}}^*_{\text{OLS}} - \boldsymbol{\delta}^* = \frac{(\widehat{\rho}_{\text{OLS}} - \rho)\boldsymbol{\delta}^* + (\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta})}{(1 - \widehat{\rho}_{\text{OLS}})}.$$

Therefore, under the homogeneity hypothesis, we have, for a large T

$$\text{Cov}\left(\sqrt{T} \widehat{\boldsymbol{\delta}}^*_{\text{OLS}} \right) = \frac{\sigma^2}{N(1-\rho)^2} \mathbf{D} \left(N^{-1} \sum_{i=1}^{N} \boldsymbol{\Psi}_i \right)^{-1} \mathbf{D}', \tag{6.65}$$

where $\mathbf{D} = (\boldsymbol{\delta}^*, \mathbf{I}_K)$ is a $K \times (K+1)$. Similarly,

$$\text{Cov}\left(\sqrt{T} \widehat{\boldsymbol{\theta}}^*_{\text{MG}} \right) = \frac{\sigma^2}{N(1-\rho)^2} \mathbf{D} \left(N^{-1} \sum_{i=1}^{N} \boldsymbol{\Psi}_i^{-1} \right) \mathbf{D}'. \tag{6.66}$$

To estimate (6.65), and (6.66), the unknown parameters σ^2, ρ, and $\boldsymbol{\theta}$ could be estimated either from pooled or mean group estimators. Using the mean group estimators, the test of the homogeneity or the long run coefficients can then be based on the following Hausman-type statistic:

$$h_{\delta^*} = NT(1 - \widehat{\rho}_{\text{MG}})^2 \left(\widehat{\boldsymbol{\delta}}^*_{\text{MG}} - \widehat{\boldsymbol{\delta}}^*_{\text{OLS}} \right)' \left(\widehat{\mathbf{D}}_{\text{MG}} \widehat{\mathbf{V}} \widehat{\mathbf{D}}'_{\text{MG}} \right)^{-1} \left(\widehat{\boldsymbol{\delta}}^*_{\text{MG}} - \widehat{\boldsymbol{\delta}}^*_{\text{OLS}} \right),$$

where $\widehat{\mathbf{D}}_{\text{MG}} = (\widehat{\boldsymbol{\delta}}^*_{\text{MG}}, \mathbf{I}_K)$, and $\widehat{\mathbf{V}}$ is given by (6.64). In general $\widehat{\mathbf{D}}_{\text{MG}} \widehat{\mathbf{V}} \widehat{\mathbf{D}}'_{\text{MG}}$ is of full rank. Under the null hypothesis, for large N and T, $h_{\delta^*} \sim \chi^2_K$.

There are two major concerns with the routine use of the Hausman procedure as a test of slope homogeneity. It could lack power for certain parameter values, as it's implicit null does not necessarily coincide with the null hypothesis of interest. Second, and more importantly, the Hausman test will not be applicable in the case of panel data models containing only strictly exogenous regressors ($\rho_i = 0$ in (6.50) for all i) or in the case of pure autoregressive models ($\boldsymbol{\beta}_i = \mathbf{0}$ in (6.50) for all i). In the former case, both estimators, $\widehat{\bar{\boldsymbol{\beta}}}_{\text{OLS}}$ and $\widehat{\bar{\boldsymbol{\beta}}}_{\text{MG}}$, are unbiased under the null and

[6] Recall that under homogeneity hypothesis $\boldsymbol{\delta}^* = \boldsymbol{\beta}/(1-\rho)$ and $\widehat{\boldsymbol{\delta}}^*_{\text{OLS}} = \widehat{\boldsymbol{\beta}}_{\text{OLS}}/(1 - \hat{\rho}_{\text{OLS}})$.

the alternative hypotheses and test will have no power. Whilst, in the case of pure autoregressive panel data models $\sqrt{NT}\left(\widehat{\bar{\rho}}_{\text{OLS}}-\bar{\rho}\right)$ and $\sqrt{NT}\left(\widehat{\bar{\rho}}_{\text{MG}}-\bar{\rho}\right)$ will be asymptotically equivalent and the asymptotic variance of $\left(\widehat{\bar{\rho}}_{\text{MG}}-\widehat{\bar{\rho}}_{\text{OLS}}\right)$ is zero under H_0.

Phillips and Sul (2003) propose a different type of Hausman test where instead of comparing two different pooled estimators of the regression coefficients (as discussed above), they propose basing the test of homogeneity on the difference between the individual estimates and a suitably defined pooled estimator. In the context of the panel regression model (6.50), their test statistic can be written as

$$G=\left(\hat{\boldsymbol{\theta}}_N-\tau_N\otimes\widehat{\bar{\boldsymbol{\theta}}}_{\text{OLS}}\right)'\hat{\Sigma}_g^{-1}\left(\hat{\boldsymbol{\theta}}_N-\tau_N\otimes\widehat{\bar{\boldsymbol{\theta}}}_{\text{OLS}}\right),$$

where $\hat{\boldsymbol{\theta}}_N=(\hat{\boldsymbol{\theta}}_1',\hat{\boldsymbol{\theta}}_2',\ldots,\hat{\boldsymbol{\theta}}_N')'$ is an $N(K+1)\times 1$ stacked vector of all the N individual estimates, τ_N is a $(N\times 1)$ vector of unity, and $\hat{\boldsymbol{\Sigma}}_g$ is a consistent estimator of $\boldsymbol{\Sigma}_g$, the asymptotic variance matrix of $\hat{\boldsymbol{\theta}}_N-\boldsymbol{\tau}_N\otimes\widehat{\bar{\boldsymbol{\theta}}}_{\text{OLS}}$, under H_0. Assuming H_0 holds and N is fixed, then $G\rightarrow_d\chi^2_{N(K+1)}$ as $T\rightarrow\infty$, so long as the $\boldsymbol{\Sigma}_g$ is a non-stochastic positive definite matrix.

As compared to the Hausman test based on $\widehat{\bar{\boldsymbol{\theta}}}_{\text{MG}}-\widehat{\bar{\boldsymbol{\theta}}}_{\text{OLS}}$, the G test is likely to be more powerful; but its use will be limited to panel data models where N is small relative to T. Also, the G test will not be valid in the case of pure dynamic models, very much for the same kind of reasons noted above in relation to the Hausman test based on $\widehat{\bar{\boldsymbol{\theta}}}_{\text{MG}}-\widehat{\bar{\boldsymbol{\theta}}}_{\text{OLS}}$. It can be shown in the case of pure autoregressive models ($\boldsymbol{\beta}_i=\mathbf{0}$ in (6.50) for all i), $\text{Rank}(\boldsymbol{\Sigma}_g)=N-1$ and $\boldsymbol{\Sigma}_g$ is non-invertible.

Swamy (1970) bases his test of slope homogeneity on the dispersion of individual estimates from a suitable pooled estimator. Swamy's test is developed for panels where N is small relative to T, but allows for cross section heteroscedasticity. Based on the Swamy's (1970) work, Pesaran and Yamagata (2008) propose standardized dispersion statistics that are asymptotically normally distributed for large N and T. Consider a modified version of Swamy's (1970) test statistic[7]

$$\tilde{S}=\sum_{i=1}^{N}\left(\widehat{\boldsymbol{\theta}}_i-\widetilde{\bar{\boldsymbol{\theta}}}_{\text{WOLS}}\right)'\frac{\mathbf{W}_i'\mathbf{W}_i}{\tilde{\sigma}_i^2}\left(\widehat{\boldsymbol{\theta}}_i-\widetilde{\bar{\boldsymbol{\theta}}}_{\text{WOLS}}\right) \tag{6.67}$$

[7] Swamy's (1970) statistic is defined by

$$\hat{S}=\sum_{i=1}^{N}\left(\widehat{\boldsymbol{\theta}}_i-\widehat{\bar{\boldsymbol{\theta}}}_{\text{WOLS}}\right)'\frac{\mathbf{W}_i'\mathbf{W}_i}{\hat{\sigma}_i^2}\left(\widehat{\boldsymbol{\theta}}_i-\widehat{\bar{\boldsymbol{\theta}}}_{\text{WOLS}}\right),$$

where

$$\widehat{\bar{\boldsymbol{\theta}}}_{\text{WOLS}}=\left(\sum_{i=1}^{N}\hat{\sigma}_i^{-2}\mathbf{W}_i'\mathbf{W}_i\right)^{-1}\sum_{i=1}^{N}\hat{\sigma}_i^{-2}\mathbf{W}_i'\mathbf{y}_i,$$

with $\hat{\sigma}_i^2=T^{-1}\left(\mathbf{y}_i-\mathbf{W}_i\widehat{\boldsymbol{\theta}}_i\right)'\left(\mathbf{y}_i-\mathbf{W}_i\widehat{\boldsymbol{\theta}}_i\right)$. Swamy shows that under H_0, $\hat{S}\rightarrow_d\chi^2_{(N-1)(K+1)}$ as $T\rightarrow\infty$ for a fixed N, and nonstochastic regressors.

where $\tilde{\sigma}_i^2$ is an estimator of σ_i^2 based on $\widehat{\widehat{\boldsymbol{\theta}}}_{\text{OLS}}$, namely

$$\tilde{\sigma}_i^2 = T^{-1}\left(\mathbf{y}_i - \mathbf{W}_i\widehat{\widehat{\boldsymbol{\theta}}}_{\text{OLS}}\right)'\left(\mathbf{y}_i - \mathbf{W}_i\widehat{\widehat{\boldsymbol{\theta}}}_{\text{OLS}}\right), \tag{6.68}$$

and $\widetilde{\widetilde{\theta}}_{\text{WOLS}}$ is the weighted pooled estimator also computed using $\tilde{\sigma}_i^2$, namely

$$\widetilde{\widetilde{\boldsymbol{\theta}}}_{\text{WOLS}} = \left(\sum_{i=1}^{N}\frac{\mathbf{W}_i'\mathbf{W}_i}{\tilde{\sigma}_i^2}\right)^{-1}\sum_{i=1}^{N}\frac{\mathbf{W}_i'\mathbf{y}_i}{\tilde{\sigma}_i^2}. \tag{6.69}$$

Suppose for the model defined by (6.50), the following relation holds:

$$N^{-1/2}\tilde{S} = N^{-1/2}\sum_{i=1}^{N} z_i + O_p\left(T^{-1}\right) + O_p\left(N^{-1/2}\right), \tag{6.70}$$

where

$$z_i = \frac{\left(T^{-1/2}\mathbf{u}_i'\mathbf{W}_i\right)\left(T^{-1}\mathbf{W}_i'\mathbf{W}_i\right)^{-1}\left(T^{-1/2}\mathbf{W}_i'\mathbf{u}_i\right)}{\mathbf{u}_i'\mathbf{u}_i/T}.$$

Since, under H_0, $z_i \to_d \chi^2_{K+1}$ as $T \to \infty$, it is reasonable to conjecture that up to order T^{-1}, $E(z_i)$ and $\upsilon_z^2 = \text{Var}(z_i)$ are given by $(K+1)$ and $2(K+1)$, respectively. Then, supposing

$$E(z_i) = (K+1) + O\left(T^{-1}\right),$$

we can write

$$N^{-1/2}\left(\frac{\tilde{S}-(K+1)}{\upsilon_z}\right) = N^{-1/2}\sum_{i=1}^{N}\left(\frac{\tilde{S}-E(z_i)}{\upsilon_z}\right) + O_p\left(\frac{\sqrt{N}}{T}\right) + O_p\left(T^{-1}\right),$$

therefore

$$\tilde{\Delta} = N^{-1/2}\left(\frac{\tilde{S}-(K+1)}{\sqrt{2(K+1)}}\right) \to_d N(0,1)$$

as N and $T \to \infty$ in no particular order, such that $\sqrt{N}/T \to 0$.

Importantly, this test is valid when the Hausman type test or G test procedure might fail to be applicable, as stated above. Moreover, this test procedure is expected to have higher power than the Hausman type test, where the latter is applicable.

6.8 A Random Coefficient Simultaneous Equation System

The generalisation of a single equation random coefficients model to a simultaneous equation system raises complicated issues of identification and estimation. To show this let us consider a system of G equations

$$\mathbf{Y}_i\mathbf{B}_i + \mathbf{X}_i\boldsymbol{\Gamma}_i = \mathbf{U}_i, \quad i = 1,\ldots,N, \tag{6.71}$$

where $\mathbf{Y}_i$ and $\mathbf{X}_i$ are the $T \times G$ and $T \times K$ matrices of endogenous and exogenous variables, respectively, $\mathbf{U}_i$ is the $T \times G$ matrices of errors, $\mathbf{B}_i$ and $\mathbf{\Gamma}_i$ are the $G \times G$ and $K \times G$ matrix of the coefficients of the endogenous variables and exogenous variables, respectively. The reduced form, then, is of the form

$$\begin{aligned} \mathbf{Y}_i &= -\mathbf{X}_i\mathbf{\Gamma}_i\mathbf{B}_i^{-1} + \mathbf{U}_i\mathbf{B}_i^{-1}, \\ &= \mathbf{X}_i\mathbf{\Pi}_i + \mathbf{V}_i, \end{aligned} \tag{6.72}$$

where

$$\mathbf{\Pi}_i = -\mathbf{\Gamma}_i\mathbf{B}_i^{-1}, \tag{6.73}$$

$$\mathbf{V}_i = \mathbf{U}_i\mathbf{B}_i^{-1}. \tag{6.74}$$

Suppose that

$$\mathbf{B}_i = \overline{\mathbf{B}} + \boldsymbol{\xi}_i, \tag{6.75}$$

$$\mathbf{\Gamma}_i = \overline{\mathbf{\Gamma}} + \boldsymbol{\alpha}_i, \tag{6.76}$$

where $\boldsymbol{\xi}_i$ and $\boldsymbol{\alpha}_i$ are $G \times G$ and $G \times K$ matrices of random variables independently distributed over i with means $\mathbf{0}$ and covariances $\mathbf{\Phi}$ and $\mathbf{\Delta}$, defined by $\mathbf{\Phi} = E[(\text{vec } \boldsymbol{\xi}_i)\ (\text{vec } \boldsymbol{\xi}_i)']$ and $\mathbf{\Delta} = E[(\text{vec } \boldsymbol{\alpha}_i)(\text{vec } \boldsymbol{\alpha}_i)']$. Then

$$\begin{aligned} E(\mathbf{\Pi}_i) &= -E[(\overline{\mathbf{\Gamma}} + \boldsymbol{\alpha}_i)(\overline{\mathbf{B}} + \boldsymbol{\xi}_i)^{-1}], \\ &\neq \overline{\mathbf{\Gamma}}\,\overline{\mathbf{B}}^{-1}. \end{aligned} \tag{6.77}$$

In other words, identification conditions of structural parameters cannot be derived by assuming that when sample size approaches infinity, $\hat{\mathbf{\Pi}}$ will converge to $\overline{\mathbf{\Gamma}}\,\overline{\mathbf{B}}^{-1}$. In fact the assumption of (6.75) raises intractable difficulties at the levels of identification and estimation.

Kelejian (1974) has studied the problem of identification under (6.75) and (6.76). His results imply that any feedback between the endogenous variables must be avoided and that identifiability and interdependence exclude each other (also see Raj and Ullah (1981)). In other words, for any one equation we may treat all the other variables as predetermined. Therefore, for ease of analysis, instead of assuming (6.75), we shall assume that

$$\mathbf{B}_i = \overline{\mathbf{B}}, \quad \forall\, i, \tag{6.78}$$

where $\overline{\mathbf{B}}$ is a non-singular matrix with fixed elements.

The combination of (6.76) and (6.78) amounts to assuming a random coefficients reduced form of (6.51), where $\mathbf{\Pi}_i = -\mathbf{\Gamma}_i\overline{\mathbf{B}}^{-1} = -(\overline{\mathbf{\Gamma}} + \boldsymbol{\alpha}_i)\overline{\mathbf{B}}^{-1}$, and

$$E(\mathbf{\Pi}_i) = -\overline{\mathbf{\Gamma}}\overline{\mathbf{B}}^{-1}, \tag{6.79}$$

$$\text{Cov}(\mathbf{\Pi}_i) = [\overline{\mathbf{B}}^{-1'} \otimes \mathbf{I}_k]\mathbf{\Delta}[\overline{\mathbf{B}}^{-1} \otimes \mathbf{I}_k] = \mathbf{\Delta}^*. \tag{6.80}$$

Assume that $\mathbf{U}_i$ are independently distributed over time but are contemporaneously correlated, then

$$\text{Cov}(\mathbf{U}_i) = E[\text{vec}(\mathbf{U}_i)\text{vec}(\mathbf{U}_i)'] = \mathbf{C}_i \otimes \mathbf{I}_T. \tag{6.81}$$

Furthermore, we assume that $\mathbf{U}_i$ and $\boldsymbol{\alpha}_i$ are mutually independent and are independent of $\mathbf{X}_i$. Then the reduced form (6.72) can be written as

$$\mathbf{Y}_i = \mathbf{X}_i\overline{\boldsymbol{\Pi}} + \mathbf{V}_i^*, \tag{6.82}$$

where $\mathbf{V}_i^* = -\mathbf{X}_i\boldsymbol{\alpha}_i\bar{\mathbf{B}}^{-1} + \mathbf{U}_i\bar{\mathbf{B}}^{-1}$ and $E(\mathbf{V}_i^*) = \mathbf{0}$,

$$\begin{aligned}\text{Cov}(\mathbf{V}_i^*) &= \bar{\mathbf{B}}^{-1'}\mathbf{C}_i\bar{\mathbf{B}}^{-1} \otimes \mathbf{I}_T + (\mathbf{I}_G \otimes \mathbf{X}_i)\Delta^*(\mathbf{I}_G \otimes \mathbf{X}_t') \\ &= \mathbf{Q}_i, \quad i = 1, \ldots, N.\end{aligned} \tag{6.83}$$

The GLS estimator of $\bar{\boldsymbol{\Pi}}$ is then equal to Balestra and Negassi (1992)

$$\text{vec}(\widehat{\overline{\boldsymbol{\Pi}}}_{\text{GLS}}) = \left[\sum_{i=1}^{N}\tilde{\mathbf{R}}_i^{-1}\right]^{-1}\left[\sum_{i=1}^{N}\tilde{\mathbf{R}}_i^{-1}\text{vec}\hat{\boldsymbol{\Pi}}_i\right], \tag{6.84}$$

where

$$\tilde{\mathbf{R}}_i = \mathbf{Q}_i \otimes (\mathbf{X}_i'\mathbf{X}_i)^{-1} + \boldsymbol{\Delta}^*, \tag{6.85}$$

$$\hat{\boldsymbol{\Pi}}_i = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{Y}_i. \tag{6.86}$$

If $\overline{\mathbf{B}}^{-1'}\mathbf{C}_i\overline{\mathbf{B}}^{-1}$ and $\boldsymbol{\Delta}^*$ are unknown, a two-step GLS procedure can be applied. In the first step, we estimate $\overline{\mathbf{B}}^{-1'}\mathbf{C}_i\overline{\mathbf{B}}^{-1}$ and $\boldsymbol{\Delta}^*$ by

$$\begin{aligned}\overline{\mathbf{B}}^{-1'}\widehat{\mathbf{C}}_i\overline{\mathbf{B}}^{-1} &= \frac{1}{T-K}\tilde{\mathbf{V}}_i^{*\prime}\tilde{\mathbf{V}}_i^*, \tilde{\mathbf{V}}_i^* = \mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\Pi}}_i, \\ \text{vec}\ (\boldsymbol{\Delta}^*) &= \frac{1}{N-1}\sum_{i=1}^{N}[\text{vec}\ (\widehat{\boldsymbol{\Pi}}_i - \overline{\widehat{\boldsymbol{\Pi}}})][\text{vec}\ (\widehat{\boldsymbol{\Pi}}_i - \overline{\widehat{\boldsymbol{\Pi}}})]', \\ \overline{\widehat{\Pi}} &= \frac{1}{N}\sum_{i=1}^{N}\widehat{\boldsymbol{\Pi}}_i.\end{aligned} \tag{6.87}$$

In the second step, we estimate $\overline{\widehat{\boldsymbol{\Pi}}}$ using (6.84) by substituting $\widehat{\tilde{\mathbf{R}}}_i$ for $\tilde{\mathbf{R}}_i$.

If our interest is in the structural form parameters $\overline{\mathbf{B}}$ and $\overline{\boldsymbol{\Gamma}}$, we can either solve for $\overline{\mathbf{B}}$ and $\overline{\boldsymbol{\Gamma}}$ from the reduced form estimate $\widehat{\overline{\boldsymbol{\Pi}}}$, or we can estimate them directly using instrumental variables method. Rewrite the first equation of the structural form in the following way,

$$\begin{aligned}\mathbf{y}_{i1} &= \mathbf{Y}_{i1}\bar{\beta}_1^* + \mathbf{X}_{i1}\gamma_{i1}^* + \mathbf{u}_{i1}, \\ &= \mathbf{Z}_{i1}\boldsymbol{\delta}_1 + \upsilon_{i1}^*, \quad i = 1, \ldots, N,\end{aligned} \tag{6.88}$$

where $\mathbf{y}_{i1}$ is the $T \times 1$ vector of the first endogenous variables and Y_{i1} is the $T \times g$ matrix of the other endogenous variables appearing in the first equation $g \leq G-1, \mathbf{X}_{i1}$ is the $T \times k^*$ matrix of included exogenous variables $k^* \leq K$, and $\bar{\beta}_1^*$ and γ_{i1}^* are $g \times 1$ and $k^* \times 1$ vectors of coefficients, respectively with $\gamma_{i1}^* = [\overline{\gamma}_{i1}^* + \boldsymbol{\alpha}_{i1}^*]$, and $\mathbf{Z}_{i1} = [\mathbf{Y}_{i1}, \mathbf{X}_{i1}], \boldsymbol{\delta}_1' = [\bar{\beta}_1^{*\prime}, \bar{\gamma}_1^{*\prime}], \mathbf{v}_{i1}^* = \mathbf{u}_{i1} + \mathbf{X}_{i1}\boldsymbol{\alpha}_{i1}^*$. Balestra and Negassi (1992) suggest the following instrumental variables estimator

$$\widehat{\boldsymbol{\delta}}_1 = \left[\sum_{i=1}^{N} \mathbf{Z}_{i1}'\mathbf{F}_i\mathbf{Z}_{i1}\right]^{-1} \left[\sum_{i=1}^{N} \mathbf{Z}_{i1}'\mathbf{F}_i\widehat{\boldsymbol{\delta}}_{i1}\right], \tag{6.89}$$

where

$$\widehat{\boldsymbol{\delta}}_{i1} = [\mathbf{Z}_{i1}'\mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{Z}_{i1}]^{-1}\mathbf{Z}_{i1}'\mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{y}_{i1}, \tag{6.90}$$

and

$$\begin{aligned} \mathbf{F}_i &= \mathbf{X}_i(\mathbf{X}_{i1}'\boldsymbol{\Delta}_1\mathbf{X}_{i1}' + \sigma_1^2\mathbf{I}_{k^*})^{-1}\mathbf{X}_i', \\ \boldsymbol{\Delta}_1 &= E\left(\boldsymbol{\alpha}_{i1}\boldsymbol{\alpha}_{i1}'\right). \end{aligned} \tag{6.91}$$

One can also derive the Bayes solutions for a simultaneous equations system of the form (6.71), (6.76) and (6.78) using a method analogous to that of Sect. 6.4. Considering one equation of (6.72) at a time, the results of sect. 4 can be applied straightforwardly. Similar results for the system of (6.72) can also be derived if the prior restrictions on $\overline{\boldsymbol{\Pi}}$ are ignored. Of course, restricted reduced form estimators can also be derived. The computation, though, can be laborious.

The results of Sect. 6.4 can also be used to derive a Bayes estimator for the structural form (6.88) based on a limited information approach. Let

$$\mathbf{Y}_{i1} = \widehat{\mathbf{Y}}_{i1} + \widehat{\mathbf{V}}_{i1}, \tag{6.92}$$

where $\widehat{\mathbf{Y}}_{i1} = \mathbf{X}_i\widehat{\boldsymbol{\Pi}}_{i1}$, and $\widehat{\boldsymbol{\Pi}}_{i1} = (\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{Y}_{i1}$. Substituting $\widehat{\mathbf{Y}}_{i1}$ for $\mathbf{Y}_{i1}$ in (6.88), we have

$$\mathbf{y}_{i1} = \widehat{\mathbf{Y}}_{i1}\overline{\boldsymbol{\beta}}_1^* + \mathbf{X}_{i1}\gamma_{i1}^* + \boldsymbol{\eta}_{i1}, \tag{6.93}$$

where $\boldsymbol{\eta}_{i1} = \mathbf{u}_{i1} + \widehat{\mathbf{V}}_{i1}\overline{\boldsymbol{\beta}}_1^*$. Conditioning on $\widehat{\boldsymbol{\Pi}}_{i1}$, we can treat $\widehat{\mathbf{Y}}_{i1}$ and $\mathbf{X}_{i1}$ as the set of exogenous variables. Equation (6.93) is of the form of the mixed fixed and random coefficients model (6.14) and the Bayes estimators of $\overline{\boldsymbol{\beta}}_1^*, \overline{\gamma}_1^*$ and γ_{i1}^* are given in Sect. 6.4 (for detail see Hsiao, Appelbe and Dineen (1992)). Of course, one should keep in mind that now the Bayes estimator is the conditional posterior mean given the estimated $\boldsymbol{\Pi}_{i1}$.

6.9 Random Coefficient Models with Cross-Section Dependence

In principle, the random coefficient model (6.14) can be easily adapted to allow for dependence across the error terms, $u_{it}, i = 1, 2, \ldots, N$. But, without plausible restrictions on the error covariances the number of unknown parameters of the model

increases at the rate of N^2, which would be manageable only when N is relatively small (typically 10 or less). To deal with the problem of cross section dependence when N is large a number of different approaches have been advanced in the literature.[8] In the case of spatial panels where a natural distance measure (or an immutable ordering of cross section units) is available the dependence is tested and modelled with "spatial lags", using techniques familiar from the time series literature. Anselin (2001) provides a recent survey of the literature on spatial econometrics. A number of studies have also used measures such as trade or capital flows to capture economic distance, as in Lee and Pesaran (1993), Conley and Topa (2002) and Pesaran, Schuermann and Weiner (2004).

But, in the absence of suitable distance measures or natural orderings of the cross section units a number of investigators have attempted to model the cross section dependence using single or multiple factor residual models where u_{it} is specified in terms of a finite number of common factors. A convenient parameterization is given by

$$u_{it} = \frac{\sigma_i}{\sqrt{1+\boldsymbol{\delta}_i'\boldsymbol{\delta}_i}}\left(\boldsymbol{\delta}_i'\mathbf{f}_t + \varepsilon_{it}\right), \tag{6.94}$$

where $\boldsymbol{\delta}_i$ is a $s \times 1$ vector of individual-specific factor loadings, $\mathbf{f}_t$ is an $s \times 1$ vector of unobserved (latent) factors, and ε_{it} is an idiosyncratic error assumed to be distributed independently across i, the unobserved factors, $\mathbf{f}_t$, and the observed regressors, $\mathbf{x}_{it}$, with mean zero and a unit variance. Since the common factors are unobserved, without loss of generality we also assume that $\mathbf{f}_t \sim (\mathbf{0}, \mathbf{I}_s)$.

Under the above set up, and conditional on a *given* set of factor loadings, the cross-correlations of the errors are given by

$$\rho_{ij} = \rho_{ji} = \frac{\boldsymbol{\delta}_i'\boldsymbol{\delta}_j}{\left(1+\boldsymbol{\delta}_i'\boldsymbol{\delta}_i\right)^{1/2}\left(1+\boldsymbol{\delta}_j'\boldsymbol{\delta}_j\right)^{1/2}}. \tag{6.95}$$

Complicated covariance structures can be accommodated by the residual factor formulation through differences across factor loadings and by using a sufficiently large number of factors. A random coefficient specification can also be assumed for the factor loadings:

$$\boldsymbol{\delta}_i = \bar{\boldsymbol{\delta}} + \boldsymbol{\zeta}_i, \tag{6.96}$$

where $\bar{\boldsymbol{\delta}}$ is a vector of fixed constants

$$\begin{aligned} E(\boldsymbol{\zeta}_i) &= \mathbf{0}, E\left(\boldsymbol{\zeta}_i\mathbf{f}_t'\right) = \mathbf{0}, \\ E\left(\boldsymbol{\zeta}_i\mathbf{x}_{it}'\right) &= \mathbf{0}, E\left(\boldsymbol{\zeta}_i\boldsymbol{\alpha}_i'\right) = \mathbf{0}, \\ E\left(\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i{}'\right) &= \begin{cases} \boldsymbol{\Delta}_\zeta, & \text{if } i = j, \\ \mathbf{0}, & \text{if } i \neq j, \end{cases} \end{aligned} \tag{6.97}$$

and Δ_ζ is a non-negative definite matrix. The average degree of cross dependence, defined by $E(\rho_{ij})$ for $i \neq j$ is governed by $\bar{\boldsymbol{\delta}}$ and the distribution of $\boldsymbol{\zeta}_i$. The average

[8] Tests of error cross section dependence in the case of large panels are proposed by Pesaran (2004).

cross section dependence will be zero if $\bar{\boldsymbol{\delta}} = \mathbf{0}$, and $\boldsymbol{\zeta}_i$ is symmetrically distributed. Typically one would expect $\bar{\boldsymbol{\delta}} \neq \mathbf{0}$.

Examples of studies that have used the residual factor structure to model cross section dependence include Holtz-Eakin, Newey, and Rosen (1988), Ahn, Lee and Schmidt (2001), Coakley, Fuertes and Smith (2005), Bai and Ng (2004), Kapetanios and Pesaran (2007), Phillips and Sul (2003), Moon and Perron (2004), and Moon, Perron and Phillips (2007) and Pesaran (2006, 2007). The studies by Holtz-Eakin et al. and Ahn et al. focus on single factor residual models and allow for time-varying individual effects in the case of panels with homogeneous slopes where T is fixed and $N \to \infty$. Phillips and Sul (2003) suggest using SURE-GLS techniques combined with median unbiased estimation in the case of first order autoregressive panels. Coakley, Fuertes and Smith (2002) propose a principal components approach which is shown by Pesaran (2006) to be consistent only when the factors and the included regressors are either asymptotically uncorrelated or are perfectly correlated. In the more general case Pesaran (2006) shows that consistent estimation of the random coefficient models with a multi-factor residual structure can be achieved (under certain regularity conditions) by augmenting the observed regressors with the cross section averages of the dependent variable and individual-specific regressors, namely

$$\bar{y}_t = \sum_{j=1}^{N} w_j y_{jt}, \text{ and } \bar{\mathbf{x}}_{it} = \sum_{j=1}^{N} w_j \mathbf{x}_{jt}, \tag{6.98}$$

for any set of weights such that

$$w_i = O\left(\frac{1}{N}\right), \sum_{i=1}^{N} |w_i| < K < \infty.$$

An obvious example of such a weighting scheme is $w_i = 1/N$.[9]

6.10 Concluding Remarks

When the included conditional variables together with the conventional variable intercept or error components (e.g. Hsiao (2003, Chap. 3)) cannot completely capture systematic differences across cross-sectional units and/or over time, and the possibility of adding additional conditional variables is not an option, either due to data unavailability or the desire to keep the model simple, there is very little alternative but to allow the slope coefficients to vary across cross-section units or over time. If we treat all these coefficients as fixed and different, there is no particular reason to pool the data, except for some efficiency gain in a Zellner's (1962) seemingly unrelated regression framework. Random coefficients models appear to be an attractive middle ground between the implausible assumption of homogeneity across

[9] Note that the non-parametric variance–covariance matrix estimator proposed in Pesaran (2006) is robust to heteroscedastic and/or serially correlated idiosyncratic errors, ε_{it}.

cross-sectional units or over time and the infeasibility of treating them all differently, in the sense of being draws from different probability distributions. Other intermediate formulations could also be considered. For example, as argued by Pesaran, Shin and Smith (1999), in the context of dynamic models it would be plausible to impose the homogeneity hypothesis on the long-run coefficients but let the short-run dynamics to vary freely across the cross-section units. In this Chapter various formulations are surveyed and their implications discussed. Our review has been largely confined to linear panel data models with stationary regressors. The analysis of random coefficient models with unit roots and cointegration is reviewed in Breitung and Pesaran (2007) in this volume. Parameter heterogeneity in non-linear panel data models poses fundamentally new problems and needs to be considered on a case-by-case basis.

Appendix A: Proof of Theorem 1

To prove part (a) of the theorem, we write (6.41) in the form of (6.19) and (6.17). Putting $\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \mathbf{C})$ and $\boldsymbol{\alpha} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_N \otimes \boldsymbol{\Delta})$ together with (6.17), the result follows.

To prove (b), we use Bayes's theorem, that is

$$p(\overline{\gamma}|\mathbf{y}) \propto p(\mathbf{y}|\overline{\gamma})p(\overline{\gamma}), \tag{6.99}$$

where $p(\mathbf{y}|\overline{\gamma})$ follows from (6.42) and $p(\overline{\gamma})$ is given by (6.39). The product on the right hand side of (6.99) is proportional to $\exp\{-\frac{1}{2}Q\}$, where Q is given by

$$\begin{aligned} Q &= (\mathbf{y} - \overline{\mathbf{Z}}\overline{\gamma})'[\mathbf{C} + \mathbf{W}(\mathbf{I}_N \otimes \Delta)\mathbf{W}']^{-1}(\mathbf{y} - \overline{\mathbf{Z}}\overline{\gamma}) \\ &= (\overline{\gamma} - \widehat{\overline{\gamma}})'\mathbf{D}^{-1}(\overline{\gamma} - \widehat{\overline{\gamma}}) + \mathbf{y}'\{\Omega^{-1} - \Omega^{-1}\overline{\mathbf{Z}}[\overline{\mathbf{Z}}'\mathbf{D}\overline{\mathbf{Z}}]^{-1}\overline{\mathbf{Z}}'\Omega^{-1}\}\mathbf{y}. \end{aligned} \tag{6.100}$$

The second term on the right hand side of (6.100) is a constant as far as the distribution of $\overline{\gamma}$ is concerned, and the remainder of the expression demonstrates the truth of (b).

To prove (c), we use the relations

$$\begin{aligned} p(\boldsymbol{\alpha}|\mathbf{y}) &= \int p(\boldsymbol{\alpha}, \overline{\gamma}|\mathbf{y})\mathrm{d}\overline{\gamma} \\ &= \int [p(\overline{\gamma}|\mathbf{y}, \boldsymbol{\alpha})d\overline{\gamma}]p(\boldsymbol{\alpha}|\mathbf{y}) \end{aligned} \tag{6.101}$$

and

$$\begin{aligned} p(\boldsymbol{\alpha}, \overline{\gamma}|\mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\alpha}, \overline{\gamma})p(\boldsymbol{\alpha}, \overline{\gamma}) \\ &= p(\mathbf{y}|\boldsymbol{\alpha}, \overline{\gamma})p(\boldsymbol{\alpha}) \cdot p(\overline{\gamma}). \end{aligned} \tag{6.102}$$

Under (6.38)–(6.40), the right hand side of (6.102) is proportional to $\exp\{-\frac{1}{2}Q^*\}$, where Q^* is given by

$$
\begin{aligned}
Q^* &= (\mathbf{y}-\overline{\mathbf{Z}}\overline{\gamma}-W\boldsymbol{\alpha})'\mathbf{C}^{-1}(\mathbf{y}-\overline{\mathbf{Z}}\overline{\gamma}-\mathbf{W}\alpha)+\boldsymbol{\alpha}'(\mathbf{I}_N\otimes\Delta^{-1})\boldsymbol{\alpha}\\
&= \mathbf{y}'\mathbf{C}^{-1}\mathbf{y}+\overline{\gamma}'\overline{\mathbf{Z}}'\mathbf{C}^{-1}\overline{\mathbf{Z}}\overline{\gamma}+\boldsymbol{\alpha}'\mathbf{W}'\mathbf{C}^{-1}\mathbf{W}\boldsymbol{\alpha}\\
&\quad -2\overline{\gamma}'\overline{\mathbf{Z}}'\mathbf{C}^{-1}\mathbf{y}-2\boldsymbol{\alpha}'\mathbf{W}'\mathbf{C}^{-1}\mathbf{y}+2\overline{\gamma}'\overline{\mathbf{Z}}'\mathbf{C}^{-1}\mathbf{W}\boldsymbol{\alpha}+\boldsymbol{\alpha}'(\mathbf{I}_N\otimes\boldsymbol{\Delta}^{-1})\boldsymbol{\alpha}\\
&= Q_1^*+Q_2^*+Q_3^*, \qquad (6.103)
\end{aligned}
$$

with

$$
\begin{aligned}
Q_1^* &= \{\overline{\gamma}-(\overline{\mathbf{Z}}'\mathbf{C}^{-1}\overline{\mathbf{Z}})^{-1}[\overline{\mathbf{Z}}'\mathbf{C}^{-1}(\mathbf{y}-\mathbf{W}\boldsymbol{\alpha})\}'(\overline{\mathbf{Z}}'\mathbf{C}^{-1}\overline{\mathbf{Z}})\\
&\quad \cdot\{\overline{\gamma}-(\overline{\mathbf{Z}}'\mathbf{C}^{-1}\overline{\mathbf{Z}})^{-1}[\overline{\mathbf{Z}}'\mathbf{C}^{-1}(\mathbf{y}-\mathbf{W}\boldsymbol{\alpha})]\}, \qquad (6.104)
\end{aligned}
$$

$$
\begin{aligned}
Q_2^* &= \{\boldsymbol{\alpha}-\widetilde{\mathbf{D}}\mathbf{W}'[\mathbf{C}^{-1}-\mathbf{C}^{-1}\overline{\mathbf{Z}}(\overline{\mathbf{Z}}'\mathbf{C}^{-1}\overline{\mathbf{Z}})^{-1}\overline{\mathbf{Z}}'\mathbf{C}^{-1}]\mathbf{y}\}'\widetilde{\mathbf{D}}^{-1}\\
&\quad \cdot\{\boldsymbol{\alpha}-\widetilde{\mathbf{D}}\mathbf{W}'[\mathbf{C}^{-1}-\mathbf{C}^{-1}\overline{\mathbf{Z}}(\overline{\mathbf{Z}}'\mathbf{C}^{-1}\overline{\mathbf{Z}})^{-1}\overline{\mathbf{Z}}'\mathbf{C}^{-1}]\mathbf{y}\} \qquad (6.105)
\end{aligned}
$$

and

$$
\begin{aligned}
Q_3^* &= \mathbf{y}'\{\mathbf{C}^{-1}-\mathbf{C}^{-1}\overline{\mathbf{Z}}(\overline{\mathbf{Z}}'\mathbf{C}^{-1}\overline{\mathbf{Z}})^{-1}\overline{\mathbf{Z}}'\mathbf{C}^{-1}-[\mathbf{C}^{-1}-\mathbf{C}^{-1}\overline{\mathbf{Z}}(\overline{\mathbf{Z}}'\mathbf{C}^{-1}\overline{\mathbf{Z}})^{-1}\overline{\mathbf{Z}}'\mathbf{C}^{-1}]\\
&\quad \cdot\mathbf{W}\widetilde{\mathbf{D}}^{-1}\mathbf{W}'[\mathbf{C}^{-1}-\mathbf{C}^{-1}\overline{\mathbf{Z}}(\overline{\mathbf{Z}}'\mathbf{C}^{-1}\overline{\mathbf{Z}})^{-1}\overline{\mathbf{Z}}'\mathbf{C}^{-1}]\}\mathbf{y}. \qquad (6.106)
\end{aligned}
$$

As far as the distribution of $p(\boldsymbol{\alpha},\overline{\gamma}|\mathbf{y})$ is concerned, Q_3^* is a constant. The conditional distribution of $\overline{\gamma}$ given $\mathbf{y}$ and $\boldsymbol{\alpha}$ is proportional to $\exp\{-\frac{1}{2}Q_1^*\}$, which integrates to 1. Therefore, the marginal distribution of $\boldsymbol{\alpha}$ given $\mathbf{y}$ is proportional to $\exp\{-\frac{1}{2}Q_2^*\}$, demonstrates (c).

Substituting (6.23)–(6.26) into (6.42) we obtain the Bayes solutions for the Swamy type random coefficients model: (i) the distribution of $\overline{\boldsymbol{\beta}}$ given $\mathbf{y}$ is $N(\widehat{\overline{\boldsymbol{\beta}}},\mathbf{D})$, and (ii) the distribution of $\boldsymbol{\alpha}$ given $\mathbf{y}$ is normal with mean

$$
\begin{aligned}
\widehat{\boldsymbol{\alpha}} &= \{\mathbf{X}'[\mathbf{C}^{-1}-\mathbf{C}^{-1}\mathbf{X}\mathbf{A}(\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}]\mathbf{X}+(\mathbf{I}_N\otimes\boldsymbol{\Delta}^{-1})\}^{-1}\\
&\quad \cdot\{\mathbf{X}'[\mathbf{C}^{-1}-\mathbf{C}^{-1}\mathbf{X}\mathbf{A}(\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}]\mathbf{y}\}\\
&= \widetilde{\mathbf{D}}\{\mathbf{X}'[\mathbf{C}^{-1}-\mathbf{C}^{-1}\mathbf{X}\mathbf{A}(\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}]\mathbf{y}\}, \qquad (6.107)
\end{aligned}
$$

and covariance

$$
\widetilde{\mathbf{D}} = \{\mathbf{X}'[\mathbf{C}^{-1}-\mathbf{C}^{-1}\mathbf{X}\mathbf{A}(\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}]\mathbf{X}+(\mathbf{I}_N\otimes\boldsymbol{\Delta}^{-1})\}^{-1}. \quad (6.108)
$$

Letting $\widetilde{\boldsymbol{\Delta}} = \mathbf{I}_N\otimes\boldsymbol{\Delta}$ and repeatedly using the identity (6.30) we can write (6.108) in the form

$$
\begin{aligned}
\widetilde{\mathbf{D}} &= [\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}+\widetilde{\boldsymbol{\Delta}}^{-1}]^{-1}\{\mathbf{I}-\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{A}[\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}+\widetilde{\boldsymbol{\Delta}}^{-1})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{A}\\
&\quad -\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{A}]^{-1}\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}[\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}+\widetilde{\boldsymbol{\Delta}}^{-1}]^{-1}\}\\
&= [\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}+\widetilde{\boldsymbol{\Delta}}^{-1}]^{-1}\{\mathbf{I}+\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{A}[\mathbf{A}'\mathbf{X}'(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})\mathbf{X}\mathbf{A}]^{-1}\\
&\quad \mathbf{A}'(\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\widetilde{\boldsymbol{\Delta}}^{-1}-\widetilde{\boldsymbol{\Delta}}^{-1})\cdot[\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}+\widetilde{\boldsymbol{\Delta}}^{-1}]^{-1}\}\\
&= [\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}+\widetilde{\boldsymbol{\Delta}}^{-1}]^{-1}+\widetilde{\boldsymbol{\Delta}}\mathbf{X}'(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{X}\mathbf{A}[\mathbf{A}'\mathbf{X}'(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{X}\mathbf{A}]^{-1}\\
&\quad \cdot\mathbf{A}'\mathbf{X}'(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{X}\widetilde{\boldsymbol{\Delta}}. \qquad (6.109)
\end{aligned}
$$

Substituting (6.109) into (6.107) we have

$$\begin{aligned}
\widetilde{\boldsymbol{\alpha}} &= [\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}+\widetilde{\boldsymbol{\Delta}}^{-1}]^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y} \\
&\quad -(\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}+\widetilde{\boldsymbol{\Delta}}^{-1})^{-1}(\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}+\widetilde{\boldsymbol{\Delta}}^{-1}-\widetilde{\boldsymbol{\Delta}}^{-1})\mathbf{A}(\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{y} \\
&\quad +\widetilde{\boldsymbol{\Delta}}\mathbf{X}'(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{X}\mathbf{A}[\mathbf{A}'\mathbf{X}'(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{X}\mathbf{A}]^{-1}\mathbf{A}'\mathbf{X}'[\mathbf{C}^{-1} \\
&\quad -(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{y}-\widetilde{\boldsymbol{\Delta}}\mathbf{X}'(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{X}\mathbf{A}[\mathbf{A}'\mathbf{X}'(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{X}\mathbf{A}]^{-1} \\
&\quad \cdot[\mathbf{I}-\mathbf{A}'\mathbf{X}(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{X}\mathbf{A}](\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{y} \\
&= (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}+\widetilde{\boldsymbol{\Delta}}^{-1})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y}-\mathbf{A}(\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{y} \\
&\quad +(\mathbf{X}'\mathbf{C}^{-1}+\widetilde{\boldsymbol{\Delta}}^{-1})^{-1}\widetilde{\boldsymbol{\Delta}}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{y} \\
&\quad -\widetilde{\boldsymbol{\Delta}}\mathbf{X}'(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{X}\mathbf{A}[\mathbf{A}'\mathbf{X}'(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{X}\mathbf{A}]^{-1}\mathbf{A}'\mathbf{X}'(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{y} \\
&\quad +\widetilde{\boldsymbol{\Delta}}\mathbf{X}'(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{X}\mathbf{A}(\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{y} \\
&= (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}+\widetilde{\boldsymbol{\Delta}}^{-1})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y}-\widetilde{\boldsymbol{\Delta}}\mathbf{X}'(\mathbf{X}\widetilde{\boldsymbol{\Delta}}\mathbf{X}'+\mathbf{C})^{-1}\mathbf{X}\mathbf{A}\widehat{\widetilde{\boldsymbol{\beta}}}. \qquad (6.110)
\end{aligned}$$

Acknowledgments We are grateful to an anonymous referee, G. Bresson, A. Pirotte, and particularly Takashi Yamagata for their careful reading of an early version and for pointing out many typos and for their good suggestions. We would also like to thank J. Breitung and Ron Smith for helpful comments.

References

Ahn, S.G., Y.H. Lee and P. Schmidt, (2001), "GMM Estimation of Linear Panel Data Models with Time-varying Individual Effects", *Journal of Econometrics*, 102, 219–255.

Amemiya, T. (1978), "A Note on a Random Coefficients Model", *International Economic Review*, 19, 793–796.

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis*, 2nd edition, New York, Wiley.

Anderson, T.W. and C. Hsiao (1981), "Estimation of Dynamic Models with Error Components", *Journal of the American Statistical Society*, 76, 598–606.

Anderson, T.W. and C. Hsiao (1982),"Formulation and Estimation of Dynamic Models Using Panel Data", *Journal of Econometrics*, 18, 47–82.

Anselin, L. (2001), "Spatial Econometrics", in B. Baltagi (ed.), *A Companion to Theoretical Econometrics*, Oxford, Blackwell.

Bai, J. and S. Ng (2004), "A Panic Attack on Unit Roots and Cointegration", *Econometrica*, 72, 1127–1177.

Balestra, P. (1996), "Introduction to Linear Models for Panel Data", in L. Mátyás and P. Sevestre (eds.), *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, New york, Springer.

Balestra, P. and S. Negassi (1992), "A Random Coefficient Simultaneous Equation System with an Application to Direct Foreign Investment by French Firms", *Empirical Economics*, 17, 205–220.

Boskin, M.J. and L.J. Lau (1990), "Post-War Economic Growth in the Group-of-Five Countries: A New Analysis", CEPR No. 217, Stanford University.

Breitung J. and M.H. Pesaran (2007), "Unit Roots and Cointegration in Panels", In this volume.
Coakley, J., A. Fuertes and R.P. Smith (2002), "A Principal Components Approach to Cross-Section Dependence in Panels", Unpublished manuscript, Birkbeck College, University of London.
Coakley, J., A. Fuertes and R.P. Smith (2005), "Unobserved Heterogeneity in Panel Time Series Models", *Computational Statistics & Data Analysis*, Forthcoming.
Conley, T.G. and G. Topa (2002), "Socio-economic Distance and Spatial Patterns in Unemployment", *Journal of Applied Econometrics*, 17, 303–327.
Cooley, T.F. and E.C. Prescott (1976) "Estimation in the Presence of Stochastic Parameter Variation", *Econometrica,* 44, 167–184.
de Finetti, B. (1964), "Foresight: Its Logical Laws. Its Subjective Sources", in J.E. Kyburg, Jr., and H.E. Smokle (eds.), *Studies in Subjective Probability*, New Yor, Wiley, 93–158.
Hausman, J.A. (1978), "Specification Tests in Econometrics", *Econometrica*, 46, 1251–1271.
Hendricks, W., R. Koenker and D.J. Poirier (1979), "Residential Demand for Electricity: An Econometric Approach", *Journal of Econometrics*, 9, 33–57.
Hildreth, C. and J.P. Houck (1968), "Some Estimators for a Linear Model with Random Coefficients", *Journal of the American Statistical Association*, 63, 584–595.
Holtz-Eakin, D, W.K. Newey and H. Rosen (1988), "Estimating Vector Autoregressions with Panel Data", *Econometrica*, 56, 1371–1395.
Hsiao, C. (1974), "Statistical Inference for a Model with Both Random Cross-Sectional and Time Effects", *International Economic Review*, 15, 12–30.
Hsiao, C. (1975), "Some Estimation Methods for a Random Coefficients Model", *Econometrica*, 43, 305–325.
Hsiao, C. (1990), "A Mixed Fixed and Random Coefficients Framework for Pooling Cross-section and Time Series Data", *Paper presented at the Third Conference on Telecommunication Demand Analysis with Dynamic Regulation*, Hilton, Head, S. Carolina.
Hsiao, C. (2003), *Analysis of Panel Data*, Economic Society monographs no. 34, 2nd ed., New York: Cambridge University Press.
Hsiao, C., T.W. Appelbe, and C.R. Dineen (1992), "A General Framework for Panel Data Models – With an Application to Canadian Customer-Dialed Long Distance Telephone Service", *Journal of Econometrics*, 59, 63–86.
Hsiao, C., D.C. Mountain, K.Y. Tsui and M.W. Luke Chan (1989), "Modeling Ontario Regional Electricity System Demand Using a Mixed Fixed and Random Coefficients Approach", *Regional Science and Urban Economics* 19, 567–587.
Hsiao, C., M.H. Pesaran and A.K. Tahmiscioglu (1999), "Bayes Estimation of Short-Run Coefficients in Dynamic Panel Data Models", in C. Hsiao, L.F. Lee, K. Lahiri and M.H. Pesaran (eds.), *Analysis of Panels and Limited Dependent Variables Models*, Cambridge: Cambridge University Press, 268–296.
Hurwicz, L. (1950), "Least Squares Bias in Time Series", in T.C. Koopman, (ed.), *Statistical Inference in Dynamic Economic Models*, New York: Wiley, 365–383.
Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl and T.C. Lee (1985), *The Theory and Practice of Econometrics*, 2nd ed., New York: Wiley.
Kapetanios, G. and M.H. Pesaran (2007), "Alternative Approaches To Estimation and Inference in Large Multifactor Panels: Small Sample Results with an Application to Modelling of Asset Return", in G. Phillips and E. Tzavalis (eds.), *The Refinement of Econometric Estimation and Test Procedures: Finite Sample and Asymptotic Analysis*, Cambridge, Cambridge University Press.
Kelejian, H.H. (1974), "Random Parameters in Simultaneous Equation Framework: Identification and Estimation", *Econometrica*, 42, 517–527.
Kiviet, J.F. and G.D.A. Phillips (1993), "Alternative Bias Approximation with Lagged-Dependent Variables", *Econometric Theory,* 9, 62–80.
Lee, K.C., and M.H. Pesaran (1993), "The Role of Sectoral Interactions in Wage Determination in the UK Economy", *The Economic Journal*, 103, 21–55.

Lindley, D.V. and A.F.M. Smith (1972), "Bayes Estimates for the Linear Model", *Journal of the Royal Statistical Society*, B. 34, 1–41.

Liu, L.M. and G.C. Tiao (1980), "Random Coefficient First-Order Autoregressive Models", *Journal of Econometrics,* 13, 305–325.

Moon, H.R. and B. Perron, (2004), "Testing for a Unit Root in Panels with Dynamic Factors", *Journal of Econometrics*, 122, 81–126.

Moon, H.R., B. Perron, and P.C.B. Phillips (2007), "Incidental Trends and the Power of Panel Unit Root Tests", in *Journal of Econometrics*, 141, 416–459 .

Murtazashvili I. and J.M. Wooldridge (2007), "Fixed Effects Instrumental Varaibles Estimation in Correlated Random Coeffcient Panel Data Models", forthcoming in *Journal of Econometrics*.

Pagan, A. (1980), "Some Identification and Estimation Results for Regression Models with Stochastically Varying Coefficients", *Journal of Econometrics*, 13, 341–364.

Pesaran, M.H. (2004), "General Diagnostic Tests for Cross Section Dependence in Panels", Cambridge Working Papers in Economics, No. 435, University of Cambridge and CESifo Working Paper Series No. 1229.

Pesaran, M.H., (2006), "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure",*Econometrica*, 74, 967–1012.

Pesaran, M.H. (2007), "A Simple Panel Unit Root Test in the Presence of Cross Section Dependence", *Journal of Applied Econometrics*, 22, 265–312.

Pesaran, M.H. and R. Smith (1995), "Estimation of Long-Run Relationships from Dynamic Heterogeneous Panels", *Journal of Econometrics*, 68, 79–114.

Pesaran, M.H., R. Smith and K.S. Im (1996), "Dynamic Linear Models for Heterogeneous Panels", Chap. 8 in Mátyás, L. and P. Sevestre (eds.), *The Econometrics of Panel Data: A Handbook of Theory with Applications*, 2nd revised edition, Doredrecht: Kluwer Academic Publications.

Pesaran, M.H.,Y. Shin and R.P. Smith, (1999), "Pooled Mean Group Estimation of Dynamic Heterogeneous Panels", *Journal of the American Statistical Association*, 94, 621–634.

Pesaran, M.H., T. Schuermann, and S.M. Weiner. (2004). "Modeling Regional Interdependencies using a Global Error-Correcting Macroeconometric Model." (with discussion), *Journal of Business and Economic Statistics* 22, 129–162 and 175–181.

Pesaran, M.H. and T. Yamagata (2008), "Testing Slope Homogeneity in Large Panels", *Journal of Econometrics*, 142, 50–93.

Phillips, P.C.B. and D. Sul (2003), "Dynamic Panel Estimation and Homogeneity Testing Under Cross Section Dependence", *Econometrics Journal*, 6, 217–259.

Raj, B. and A. Ullah (1981), *Econometrics: A Varying Coefficients Approach*, London, Croom Helm.

Rosenberg, B. (1972), "The Estimation of Stationary Stochastic Regression Parameters Reexamined", *Journal of the American Statistical Association*, 67, 650–654.

Rosenberg, B. (1973), "The Analysis of a Cross-Section of Time Series by Stochastically Convergent Parameter Regression", *Annals of Economic and Social Measurement*, 2, 399–428.

Singh, B., A.L. Nagar, N.K. Choudhry, and B. Raj (1976), "On the Estimation of Structural Changes: A Generalisation of the Random Coefficients Regression Model", *International Economic Review*, 17, 340–361.

Swamy, P.A.V.B. (1970), "Efficient Inference in a Random Coefficient Regression Model", *Econometrica*, 38, 311–323.

Swamy, P.A.V.B. and P.A. Tinsley (1977), "Linear Prediction and Estimation Methods for Regression Models with Stationary Stochastic Coefficients", *Federal Reserve Board Division of Research and Statistics*, Special Studies Paper No. 78, Washington, D.C.

Wachter, M.L. (1976), "The Changing Cyclical Responsiveness of Wage Inflation", *Brookings Paper on Economic Activity*, 115–168.

Zellner, A. (1962), "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias", *Journal of the American Statistical Association*, 57, 348–368.

Chapter 7
Parametric Binary Choice Models

Michael Lechner, Stéfan Lollivier and Thierry Magnac

7.1 Introduction

Binary dependent data are a common feature in many areas of empirical economics as, for example, in transportation choice, the analysis of unemployment, labor supply, schooling decisions, fertility decisions, innovation behaviour of firms, etc. As panel data is increasingly available, the demand for panel data models coping with binary dependent variables is also increasing. Also, dramatic increases in computer capacity have greatly enhanced our ability to estimate a new generation of models. The second volume of this handbook contains several applications based on this type of dependent variable and we will therefore limit this chapter to the exposition of econometric models and methods.

There is a long history of binary choice models applied to panel data which can for example be found in Arellano and Honore (2001), Baltagi (2000), Hsiao (1992, 1996, 2003), Lee (2002) or Sevestre (2002) as well as in chapters of econometrics textbooks as for instance Greene (2003) or Wooldridge (2002). Some of these books and chapters do not devote much space to the binary choice model. Here, in view of other chapters in this handbook that address related nonlinear models (qualitative, truncated or censored variables, nonparametric models, etc.), we focus on the parametric binary choice model and some of its semiparametric extensions. The binary choice model provides a convenient benchmark case, from which many results can be generalised to limited dependent variable models such as multinomial discrete choices (Train, 2002), transition models in continuous time (Kamionka, 1998) or to structural dynamic discrete choice models that are not studied here.

Michael Lechner
Swiss Institute for Empirical Economic Research (SEW), University of St. Gallen, Varnbühlstr. 14, CH-9000 St. Gallen, Switzerland, e-mail: Michael.Lechner@unisg.ch

Stéfan Lollivier
INSEE, 18 boulevard Adolphe Pinard, F-75014 Paris, France, e-mail: stefan.lollivier@insee.fr

Thierry Magnac
Université de Toulouse 1, Toulouse School of Economics, Manufacture des Tabacs, 21, Allée de Brienne, 31000 Toulouse, France, e-mail: magnac@cict.fr

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

We tried to be more comprehensive than the papers and chapters mentioned and we provide an introduction into the many issues that arise in such models. We also try not only to provide an overview of different models and estimators but also to make sure that the technical level of this chapter is such that it can easily be understood by the applied econometrician. For all technical details, the reader is referred to the specific papers.

Before we discuss different versions of the binary choice panel data models, define first the notation for the data generating process underlying the prototypical binary choice panel model:

$$y_{it} = \mathbf{1}\{y_{it}^* > 0\} \text{ for any } i = 1, \ldots, N \text{ and } t = 1, \ldots, T,$$

where $\mathbf{1}\{.\}$ is the indicator of the event between brackets and where the latent dependent variables y_{it}^* are written as:

$$y_{it}^* = X_{it}\beta + \varepsilon_{it},$$

where β denotes a vector of parameters, X_{it} is a $1 \times K$ vector of explanatory variables and error terms ε_{it} stand for other unobserved variables. Stacking the T observations of individual i,

$$Y_i^* = X_i\beta + \varepsilon_i \,,$$

where $Y_i^* = (y_{i1}^*, ., y_{iT}^*)$ is the vector of latent variables, $X_i = (X_{i1}, ., X_{iT})$ is the $T \times K$ matrix of explanatory variables and $\varepsilon_i = (\varepsilon_{i1}, ., \varepsilon_{iT})$ is the $T \times 1$ vector of errors.

We focus on the estimation of parameter β and of parameters entering the distribution function of ε_{it}. We do not discuss assumptions under which such parameters can be used to compute other parameters, such as causal effects (Angrist, 2001). We also consider balanced panel data for ease of notation although the general case of unbalanced panel is generally not much more difficult if the data is missing at random (see Chap. 12).

As usual in econometrics we impose particular assumptions at the level of the latent model to generate the different versions of the observable model to be discussed in the sections of this chapter. These assumptions concern the correlation of the error terms over time as well as the correlation between the error terms and the explanatory variables. The properties for various conditional expectations of the observable binary dependent variable are then derived. We assume that the observations are obtained by independent draws in the population of statistical units 'i', also called individuals in this chapter. Working samples that we have in mind are much larger in dimension N than in dimension T and in most cases we consider asymptotics in N holding T fixed although we report on some recent work on large T approximations. Time effects can then be treated in a determistic way. In this chapter we frequently state our results for an important special case, the panel probit model where error terms ε_i are assumed to be normally distributed.

In Sect. 7.2 of this chapter we discuss different versions of the static random effects model when the explanatory variables are strictly exogenous. Depending on the autocorrelation structure of the errors different estimators are available and

we detail their attractiveness in each situation by trading-off their efficiency and robustness with respect to misspecification. Section 7.3 considers the static model when a time invariant unobservable variable is correlated with the time varying explanatory variables. The non linearity of binary choice models makes it pretty hard to eliminate individual fixed effects in likelihood functions and moment conditions, because the usual 'differencing out trick' of the linear model does not work except in special cases. Imposing quite restrictive assumptions is the price to pay to estimate consistently parameters of interest. Finally, Sect. 7.4 addresses the important issue of structural dynamics for fixed and random effects, in other words cases when the explanatory variables include lagged endogenous variables or are weakly exogenous only.

7.2 Random Effects Models Under Strict Exogeneity

In this section we set up the simplest models and notations that will be used in the rest of the chapter. We consider in this chapter that random effects models defined as in Arellano and Honore (2001) as models where errors in the latent model are independent of the explanatory variables.[1] This assumption does not hold with respect to the explanatory variables in the current period only but also in all past and future periods so that explanatory variables are also considered in this section to be strictly exogenous in the sense that:

$$F_{\varepsilon_t}(\varepsilon_{it}|X_i) = F_{\varepsilon_t}(\varepsilon_{it}) \,, \tag{7.1}$$

where $F_{\varepsilon_t}(\varepsilon_{it})$ denotes the marginal distribution function of the error term in period t. When errors are not independent over time, it will also at times be useful to impose a stronger condition on the joint distribution of the T errors terms over time, denoted $F_{\varepsilon}^{(T)}(\cdot)$:

$$F_{\varepsilon}^{(T)}(\varepsilon_i|X_i) = F_{\varepsilon}^{(T)}(\varepsilon_i) \,. \tag{7.2}$$

Note that as in binary choice models in cross-sections, marginal choice probabilities can be expressed in terms of the parameters of the latent model:

$$\begin{aligned} P(y_{it}=1|X_i) &= E(y_{it}=1|X_i) \\ &= E(y_{it}=1|X_{it}=x_{it}) = 1 - F_{\varepsilon_t}(-X_{it}\beta) \,. \end{aligned} \tag{7.3}$$

It also emphasizes that the expectation of a Bernoulli variable completely describes its distribution.

[1] One needs to assume independence between errors and regressors instead of assuming that correlations are equal to zero because of the non-linearity of the conditional expectation of the dependent variable **with respect to individual effects**.

We already said that we would consider random samples only. Individual observations are independent and if θ generically denote all unknown parameters including those of the distribution function of errors, the sample likelihood function is the product of individual likelihood functions:

$$L(\theta) = \prod_{i=1}^{N} L_i(Y_i|X_i;\theta)$$

where $Y_i = (y_{i1}, ., y_{iT})$ is the vector of binary observations.

7.2.1 Errors are Independent Over Time

When errors are independent over time, the panel model collapses to a cross-sectional model with NT independent observations and the maximum likelihood estimator is the standard estimator of choice. The likelihood function for one observation is given by:

$$L_i(Y_i|X_i;\theta) = \prod_{t=1}^{T} [1 - F_{\varepsilon_t}(-X_{it}\beta)]^{y_{it}} F_{\varepsilon_t}(-X_{it}\beta)^{(1-y_{it})} . \tag{7.4}$$

Later it will be pointed out that even if true errors are not independent over time, nevertheless the pseudo-maximum likelihood estimator (incorrectly) based on independence – the so called 'pooled estimator' – has attractive properties (Robinson, 1982).

Let $\Phi(\cdot)$ denote the cumulative distribution function (cdf) of the univariate zero mean unit variance normal distribution, we obtain the following log-likelihood function for the probit model:

$$\begin{aligned} &L_i(Y_i|X_i;\beta,\sigma_2,\ldots,\sigma_T;\sigma_1 = 1) \\ &\quad = \sum_{t=1}^{T} y_{it} \ln \Phi\left(\frac{X_{it}\beta}{\sigma_t}\right) + (1 - y_{it}) \ln[1 - \Phi\left(\frac{X_{it}\beta}{\sigma_t}\right)]. \end{aligned}$$

Note that to identify the scale of the parameters, the standard error of the error term in the first period is normalised to 1 ($\sigma_1 = 1$). If all coefficients are allowed to vary over time in an unrestricted way, then more variances have to be normalised.[2] In many applications however, the variance of the error is kept constant over time ($\sigma_t = 1$). For notational convenience this assumption will be maintained in the remainder of the chapter.

[2] See for example the discussion in Chamberlain (1984).

7.2.2 One Factor Error Terms

7.2.2.1 The Model

Probably the most immediate generalisation of the assumption of independent errors over time is a one-factor structure where all error terms are decomposed into two different independent components. One is constant over time (u_i) and is called the individual effect, the other one being time variable (v_{it}), but identically and independently distributed (iid) over time and individuals. Thus, we assume that for $i = 1, \ldots, N$ and $t = 1, \ldots, T$:

$$\varepsilon_{it} = u_i + v_{it}, \quad F_v^{(T)}(v_{i1}, \ldots, v_{iT}|X_i) = \prod_{t=1}^{T} F_{v_t}(v_{it}) ;$$

$$F_{u,v}^{(T)}(u_i, v_{i1}, \ldots, v_{iT}|X_i) = F_u(u_i) \prod_{t=1}^{T} F_{v_t}(v_{it}) .$$

The individual effect, u_i, can be interpreted as describing the influence of time-independent variables which are omitted from the model and that are independent of the explanatory variables. Note that the one-factor decomposition is quite strong in terms of its time series properties, because the correlation between the error terms of the latent model does not die out when the time distance between them is increased.

To achieve identification, restrictions need to be imposed on the variances of each error component which are denoted σ_v^2 and σ_u^2. For example, variance σ_v^2 can be assumed to be equal to a given value (to 1 in the normal case), or one can consider the restriction that the variance of the sum of error terms is equal to 1 ($\sigma_u^2 + \sigma_v^2 = 1$). It simplifies the comparison with cross section estimations. In this section, we do not restrict σ_u and σ_v for ease of notation though such a restriction should be imposed at the estimation stage.

7.2.2.2 Maximum Likelihood Estimation

The computation of the log-likelihood function is difficult when errors are not independent over time or have not a one-factor structure since the individual likelihood contribution is defined as an integral with respect to a T dimensional distribution function. Assumptions of independence or one-factor structure simplify the computation of the likelihood function (Butler and Moffitt, 1982).

The idea is the following. For a given value of u_i, the model is a standard binary choice model as the remaining error terms v_{it} are independent between dates and individuals. Conditional on u_i, the likelihood function of individual i is thus:

$$L_i(Y_i|X_i, u_i; \theta) = \prod_{t=1}^{T} \left[[1 - F_v(-X_{it}\beta - u_i)]^{y_{it}} [F_v(-X_{it}\beta - u_i)]^{1-y_{it}} \right]$$

The unconditional likelihood function is derived by integration:

$$L_i(Y_i|X_i;\theta) = \int_{-\infty}^{+\infty} L_i(Y_i|X_i,u_i;\theta) f_u(u_i)\mathrm{d}u_i \ . \tag{7.5}$$

The computation of the likelihood function thus requires simple integrations only. Moreover, different parametric distribution functions for u_i and v_{it} can be specified in this 'integrating out' approach. For instance, the marginal distribution functions of the two error components can be different as in the case with a normal random effect and logistic iid random error.[3] Also note that the random effect may be modelled in a flexible way. For example Heckman and Singer (1984), Mroz (1999), and many others suggested the modeling framework where the support of individual effects of u_i is discrete so that the cumulative distribution function of u_i is a step function. Geweke and Keane (2001) also suggest mixtures of normal distribution functions.

For the special case of a T normal variate error, u_i, the log-likelihood of the resulting probit model is given by:

$$\begin{aligned} L_i(Y_i|X_i;\theta) &= \\ &= \int_{-\infty}^{+\infty} \left\{ \prod_{t=1}^{T} \left[\Phi\left(\frac{X_{it}\beta + \sigma_u u_i}{\sigma_v} \right) \right]^{y_{it}} \left[1 - \Phi\left(\frac{X_{it}\beta + \sigma_u u_i}{\sigma_v} \right) \right]^{1-y_{it}} \right\} \phi(u_i)\mathrm{d}u_i, \end{aligned} \tag{7.6}$$

where $\phi(\cdot)$ denotes the density function of the standard normal distribution. In this case, the most usual identification restriction is $\sigma_u^2 + \sigma_v^2 = 1$, so that the disturbances can be written as:

$$\varepsilon_{it} = \gamma u_i + \sqrt{1-\gamma^2} v_{it} \ ,$$

where u_i and v_{it} are univariate normal, $N(0,1)$, and $\gamma > 0$. Parameter γ^2 is the share of the variance of the error term due to individual effects.

The computation of the likelihood function is a well-known problem in mathematics and is performed using gaussian quadrature. The most efficient method of computation that leads to the so called 'random effects probit estimator' uses the Hermite integration formula (Butler and Moffitt, 1982). See also the paper by Guilkey and Murphy (1993) for more details on this model and estimator as well as Lee (2000) for more discussion about the numerical algorithm.

Finally, Robinson (1982) and Avery, Hansen and Hotz (1983) show that the pooled estimator is an alternative to the previous method. The pooled estimator is the pseudo-maximum likelihood estimator where it is incorrectly assumed that errors are independent over time. As a pseudo likelihood estimator, it is consistent though inefficient. Note that the standard errors of estimated parameters are to be computed using pseudo-likelihood theory (Gouriéroux, Monfort and Trognon, 1984).

[3] As it can be found in STATA for instance.

7.2.3 General Error Structures

Obviously, the autocorrelation structure implied by the one factor-structure is very restrictive. Correlations do not depend on the distance between periods t and t'. The general model that uses only the restrictions implied by (7.1) and (7.2) poses, however severe computational problems. Computing the maximum likelihood estimator requires high dimensional numerical integration. For example, Gaussian quadrature methods for the normal model do not work in practice when the dimension of integration is larger than four.

There are two ways out of these computational problems. First, instead of computing the exact maximum likelihood estimator, we can use simulation methods and approximate the ML estimator by simulated maximum likelihood (SML). It retains asymptotic efficiency under some conditions that will be stated later on (e.g. Hajivassiliou, McFadden and Ruud (1996)). In particular, SML methods require that the number of simulations tends to infinity to obtain consistent estimators. As an alternative there are estimators which are more robust to misspecifications in the serial correlation structure but which are inefficient because they are either based on misspecified likelihood functions (pseudo-likelihood) or on moment conditions that do not depend on the correlation structure of the error terms (GMM, e.g. Avery, Hansen and Hotz (1983), Breitung and Lechner (1997), Bertschek and Lechner (1998) and Inkmann (2000)). Concerning pseudo-ML estimation, we already noted that the pooled probit estimator is consistent irrespective of the error structure. Such a consistency proof is however not available for the one-factor random effects probit estimator.

Define the following set function:

$$D(Y_i) = \left\{ Y_i^* \in \mathbb{R}^T \text{ such that } \begin{array}{c} 0 \leq y_{it}^* < +\infty \text{ if } y_{it} = 1 \\ -\infty < y_{it}^* < 0 \text{ if } y_{it} = 0 \end{array} \right\} \tag{7.7}$$

The contribution of observation i to the likelihood is:

$$L_i(Y_i \,|X_i;\theta) = E\left[\mathbf{1}\left\{Y_i^* \in D(Y_i)\right\}\right] \tag{7.8}$$

In probit models, ε_i is distributed as multivariate normal $N(0,\Omega)$, Ω being a $T \times T$ variance–covariance matrix. The likelihood function is:

$$L_i(Y_i|X_i;\theta) = \int_{D(Y_i)} \phi^{(T)}(Y_i^* - X_i\beta,\Omega)\mathrm{d}Y_i^*,$$

where $\phi^{(T)}(\cdot)$ denotes the density of the T-variate normal distribution.

In the general case, the covariance matrix of the errors Ω is unrestricted (except for identification purposes, see above). It is very frequent however to restrict its structure to reduce the number of parameters to be estimated. The reason for doing so is computation time, stability of convergence, occurrence of local extrema and the difficulties to pin down (locally identify) the matrix of correlations when the sample size is not very large. In many applications the random effects model discussed in

the previous section is generalised by allowing for an AR(1) process in the time variant error component (v_{it}). Other more general structures however are feasible as well if there are enough data.

We will see below how to use simulation to approximate the likelihood function by using Simulated Maximum Likelihood (SML). Another popular estimation method consist in using conditional moments directly. They are derived from the true likelihood function and are approximated by simulation (Method of Simulated Moments or MSM). McFadden (1989) proposed to consider all possible sequences of binary variables over T periods, Y_ω, where ω runs from 1 to 2^T. Choice indicators are defined as $d_{i\omega} = 1$ if i chooses sequence ω and is equal to 0 otherwise. A moment estimator solves the empirical counterpart of the moment condition:

$$E\left[\sum_{\omega=1}^{2^T} W_{i\omega}\left[d_{i\omega} - P_{i\omega}(\theta)\right]\right] = 0\,, \tag{7.9}$$

where $P_{i\omega}(\theta) = L_i(Y_\omega \mid X_i;\theta)$ is the probability of sequence ω (i.e. such that $Y_i = Y_\omega$). The optimal matrix of instruments $W_{i\omega}$ in the moment condition is:

$$W^*_{i\omega} = \left.\frac{\partial \log[P_{i\omega}(\theta)]}{\partial\theta}\right|_{\theta=\theta_0}\,,$$

where parameter θ_0 is the true value of θ. In practice, any consistent estimator is a good choice to approximate parameter θ_0. The first of a two-step GMM procedure using the moment conditions above and identity weights can lead to such a consistent estimate. It is then plugged in the expression for $W^*_{i\omega}$ at the second step.

Even if T is moderately large however, the number of sequences ω is geometric in $T(2^T)$ and functions $P_{i\omega}(\theta)$ can be very small. What proposes Keane (1994) is to replace in (7.9) unconditional probabilities by conditional probabilities:

$$E\left[\sum_{t=1}^{T}\sum_{j=0}^{1} \tilde{W}_{it\,j}\left(d_{it\,j} - P_{it\,j}(\theta)\right)\right] = 0\,,$$

where $d_{it\,j} = 1$ if and only if $y_{it} = j$ and where:

$$\begin{aligned} P_{it\,j}(\theta) &= P(y_{it} = j \mid y_{i1},.,y_{it-1},X_i;\theta) \\ &= \frac{P(y_{it} = j, y_{i1},.,y_{it-1} \mid X_i;\theta)}{P(y_{i1},.,y_{it-1} \mid X_i;\theta)} \end{aligned}$$

is the conditional probability of choice j conditional on observed lagged choices.

Finally, maximising the expectation of the log-likelihood function $E\,\log[L_i(Y_i \mid X_i;\theta)]$ is equivalent to solving the following system of score equations with respect to θ:

$$E\left[S_i(\theta)\right] = 0\,,$$

where $S_i(\theta) = \frac{\partial \log[L_i(Y_i|X_i;\theta)]}{\partial \theta}$ is the score function for individual i. It can be shown that, in most limited dependent variable models (Hajivassiliou and McFadden, 1998):

$$\frac{\partial}{\partial \theta} L_i(Y_i \mid X_i;\theta) = E\left[g_i(Y_i^* - X_i\beta)\mathbf{1}\{Y_i^* \in D(Y_i)\}\right]]$$

where:

$$g_i(u) = \begin{bmatrix} X_i'\Omega^{-1}u \\ \Omega^{-1}(uu' - \Omega)\Omega^{-1}/2 \end{bmatrix}$$

The score function can then be written as a conditional expectation:

$$S_i(\theta) = E\left[g_i(Y_i^* - X_i\beta)\,|Y_i^* \in D(Y_i)\right] \tag{7.10}$$

which opens up the possibility of computing the scores by simulations (Method of Simulated Scores, MSS, Hajivassiliou and McFadden, 1998).

7.2.4 Simulation Methods

Simulation methods (SML, MSM, MSS) based on the criteria established in the previous section consist in computing the expectation of a function of T random variates. The exact values of these high dimensional integrals are too difficult to compute and these expectations are approximated by sums of random draws using laws of large numbers:

$$\frac{1}{H}\sum_{h=1}^{H} f(\varepsilon_h) \underset{H\to\infty}{\overset{P}{\to}} Ef(\varepsilon)$$

when ε_h is a random draw from a distribution. In the case of panel probit models, it is a multivariate normal distribution function, $N(0,\Omega)$.

It is not the purpose of this chapter to review the general theory of simulation (see Gouriéroux and Monfort (1996) and Geweke and Keane (2001)). We review the properties of such methods in panel probit models only to which we add a brief explanation of Gibbs resampling methods which borrow their principle from Bayesian techniques.

7.2.4.1 The Comparison Between SML, MSM, MSS in Probit Models

The naive SML function is for instance:

$$\frac{1}{H}\sum_{h=1}^{H} I\{Y_i^* \in D(Y_i)\}$$

where $I[Y_i^* \in D(Y_i)]$ is a simulator. It is not continuous with respect to the parameter of interest however and this simulation method is not recommendable.

What is recommended is to use a smooth simulator which is differentiable with respect to the parameter of interest. The Monte Carlo evidence that the Geweke–Hajivassiliou–Keane (GHK) simulator is the best one in multivariate probit models seems overwhelming (see Geweke and Keane (2001) and Hajivassiliou, McFadden and Ruud (1996), for a presentation).

The asymptotic conditions concerning the number of draws (H) and leading to consistency, absence of asymptotic bias and asymptotic normality are more or less restrictive according to each method, SML, MSM or MSS (Gouriéroux and Monfort, 1993). The method of simulated moments (MSM) yields consistent, asymptotically unbiased and normally distributed estimators as $N \to \infty$ when H is fixed because the moment condition (7.9) is linear in the simulated expression (or the expectation). In Keane's (1994) version of MSM where conditional probabilities are computed by taking ratios, the estimator is only consistent when the number of draws tends to infinity. Similarly, because a logarithmic transformation is taken, SML is not consistent when H is fixed. Consistency is obtained when H grows at any rate towards infinity (Lee, 1992). Furthermore, a sufficient condition to obtain asymptotically unbiased, asymptotically normal and efficient estimates is $\sqrt{N}/H \to 0$ as $N \to \infty$ (Lee, 1992; Gouriéroux and Monfort, 1993).

It is the reason why some authors prefer MSM to SML. As already said, MSM however requires the computation of the probabilities of all the potential paths with longitudinal data although the less intensive method proposed by Keane (1994) seems to work well in panel probit models (Geweke, Keane and Runkle, 1997). The computation becomes cumbersome when the number of periods is large and there is evidence that small sample biases in MSM are much larger than the simulation bias (Geweke and Keane, 2001). Lee (1995) proposed procedures to correct asymptotic biases though results are far from impressive (Lee, 1997; Magnac, 2000). The GHK simulator is an accurate simulator though it may require a large number of draws to be close to competitors such as Monte Carlo Markov Chains (MCMC) methods (Geweke, Keane and Runkle, 1997). There seems to be a general consensus between authors about the deterioration of all estimators when the amount of serial correlation increases.

Another way to obtain consistent estimators for fixed H is the method of simulated scores (MSS) if the simulator is unbiased. It seems that it is simpler than MSM because it implicitly solves the search for optimal instruments. Hajivassiliou and McFadden (1998) proposes an acceptance–rejection algorithm consisting in rejecting the draw if the condition in (7.10) is not verified. The simulator is not smooth however and as already said a smooth simulator seems to be a guarantee of stability and success for an estimation method. Moreover, in particular when T exceeds four or five, it is possible for some individuals that the acceptance condition is so strong that no draw is accepted. Other methods consist in considering algorithms either based on GHK simulations of the score or on Gibbs resampling. Formulas and an evaluation are given in Hajivassiliou, McFadden and Ruud (1996).[4]

[4] Hajivassiliou and McFadden (1998) first propose to simulate the numerator and the denominator separately. Of course, this method does not lead to unbiased simulation because the ratio is not linear but, still, as simulators are asymptotically unbiased, those MSS estimators are consistent whenever H tends to infinity. The authors furthermore argue that using the same random draws

7.2.4.2 Gibbs Sampling and Data Augmentation

It is possible however to avoid maximisation by applying Gibbs sampling techniques and data augmentation in multiperiod probit models (Geweke, Keane and Runkle, 1997; Chib and Greenberg, 1998; Chib, 2001). Though the original setting of Monte Carlo Markov Chains (MCMC) is Bayesian, it can be applied to classical settings as shown by Geweke, Keane and Runkle (1997). The posterior density function of parameter θ given the data $(Y,X) = \{(Y_i,X_i); i = 1,.,n\}$ can indeed be used to compute posterior means and variance–covariance matrices to be used as classical estimators and their variance–covariance matrices.

To compute the posterior density $p(\theta \mid Y,X)$, we rely on two tools. One is the Metropolis–Hastings algorithm which allows for drawing samples in any (well behaved) multivariate density function, the other is Gibbs resampling which allows to draw in the conditional densities instead of the joint density function.

In the case of panel probit models, it runs as follows. First, let us 'augment' the data by introducing the unknown latent variables $Y_i^* = X_i\beta + \varepsilon$ in order to draw from the posterior density $p(\theta, Y^* \mid Y,X)$ instead of the original density function. The reason is that it will be much easier to sample into density functions conditional on the missing latent variables. Second, parameter θ is decomposed into different blocks $(\theta_1,.,\theta_J)$ according to the different types of parameters in β or in Ω the variance–covariance matrix.[5]

Let us choose some initial values for θ, say $\theta^{(0)}$ and proceed as follows. Draw Y^* in the distribution function $p(Y^* \mid \theta^{(0)},Y,X)$ – it is a multivariate truncated normal density function – in a very similar way to the GHK simulator. Then draw a new value for the first block θ_1 in θ, i.e., from $p(\theta_1 \mid Y^*, \theta_{-1}^{(0)}, Y, X)$ where $\theta_{-1}^{(0)}$ is constructed from parameter $\theta^{(0)}$ by omitting $\theta_1^{(0)}$. Denote this draw $\theta_1^{(1)}$. Do similar steps for all blocks $j = 2,\ldots,J$, using the updated parameters, until a new value $\theta^{(1)}$ is completed. Details of each step are given in Chib and Greenberg (1998). Repeat the whole step M times – M depends on the structure of the problem (Chib, 2001). Trim the beginning of the sample $\{\theta^{(0)},\ldots,\theta^{(m)}\}$, the first 200 observations say. Then, the empirical density function of $\{\theta^{(m+1)},\ldots,\theta^{(M)}\}$ is $p(\theta \mid Y,X)$. Once again, this method is computer intensive with large samples and many dates. It is however a close competitor to SML and MSS (Geweke and Keane, 2001).

7.2.4.3 Using Marginal Moments and GMM

Instead of working with the joint distribution function, the model defined by (7.8) implies the following moment conditions about the marginal period-by-period distribution functions.[6]

---

for the denominator and the numerator decreases the noise. The other method based on Gibbs resampling seems expensive in terms of computations using large samples though it is asymptocally unbiased as soon as H tends to infinity faster than $\log(N)$.

[5] See Chib and Greenberg (1998) to assess how to do the division into blocks according to the identifying or other restrictions on parameter β or on matrix Ω.

[6] The following section heavily draws from Bertschek and Lechner (1998).

$$
\begin{aligned}
&E[M(Y,X;\beta_0)|X] = 0,\\
&M(Y,X;\beta) = [m_1(Y_1,X;\beta),\ldots,m_t(Y_t,X;\beta),\ldots,m_T(Y_T,X;\beta)]',\\
&m_t(Y_t,X;\beta) = Y_t - [1 - F(-X_t\beta)]\,.
\end{aligned}
\tag{7.11}
$$

For the probit model the last expression specialises to $m_t(Y_t,X_t;\beta) = Y_t - \Phi(X_t\beta)$. Although the conditional moment estimator (CME) based on these marginal moments will be less efficient than full information maximum likelihood (FIML), these moment estimators have the clear advantage that fast and accurate approximation algorithms are available and that they do not depend on the off-diagonal elements of the covariance matrix of the error terms. Thus, these nuisance parameters need not be estimated to obtain consistent estimates of the scaled slope parameters of the latent model. At least, these estimators yields interesting initial conditions and previous methods can be used to increase efficiency.

As in the full information case, there remains the issue of specifying the instrument matrix. First, let us consider a way to use these marginal moments under our current set of assumptions in the asymptotically efficient way. Optimal instruments are given by:

$$A^*(X_i,\theta_0) = D(X_i,\theta_0)'\Omega(X_i,\theta_0)^{-1}\,;$$

$$D(X_i,\theta) = E\frac{\partial M(Y,X_i,\theta)}{\partial\theta}|X = X_i\,; \tag{7.12}$$

$$\Omega(X_i,\theta) = E[M(Y,X_i,\theta)M(Y,X_i,\theta)']|X = X_i\,. \tag{7.13}$$

For the special case of the probit model under strict exogeneity the two other elements of (7.13) have the following form:

$$D_{it}(X_{it};\beta_0) = -\phi(X_{it}\beta_0)X_{it}$$

$$\omega_{its}(X_{it},\beta_0) = [E(Y_t - \Phi_{it})(Y_s - \Phi_{is})|X = X_i] \tag{7.14}$$

$$= \begin{cases} \Phi_{it}(1-\Phi_{it}) \text{ if } t = s \\ \Phi^{(2)}_{its} - \Phi_{it}\Phi_{is} \text{ if } t \neq s \end{cases} \tag{7.15}$$

where $\Phi_{it} = \Phi(X_{it}\beta_0)$ and $\Phi^{(2)}_{its} = \Phi^{(2)}(X_{it}\beta_0, X_{is}\beta_0, \rho_{ts})$ denotes the cdf of the bivariate normal distribution with correlation coefficient ρ_{ts}. The estimation of the optimal instruments is cumbersome because they vary with the regressors in a nonlinear way and depend on the correlation coefficients.

There are several different ways to obtain consistent estimates of the optimal instruments. Bertschek and Lechner (1998) propose to estimate the conditional matrix nonparametrically. They focus on the k-nearest neighbour (k-NN) approach to estimate $\Omega(X_i)$, because of its simplicity. k-NN averages locally over functions of the data of those observations belonging to the k-nearest neighbours. Under regularity conditions (Newey, 1993), this gives consistent estimates of $\Omega(X_i)$ evaluated at $\tilde{\beta}_N$ and denoted by $\tilde{\Omega}(X_i)$ for each observation 'i' without the need for estimating ρ_{ts}. Thus, an element of $\Omega(X_i)$ is estimated by:

$$\tilde{\omega}_{its}(X_i) = \sum_{j=1}^{N} w_{ijts} m_t(y_{jt}, X_{jt}; \tilde{\beta}_N) m_s(y_{it}, X_{it}; \tilde{\beta}_N) \,, \tag{7.16}$$

where w_{ijts} represents a weight function. This does not involve an integral over a bivariate distribution. For more details on different variants of the estimator and how to implement it, the reader is referred to Bertschek and Lechner (1998). In their Monte Carlo study it appeared that optimal (nonparametric) Conditional Moment estimators based on moments rescaled to have a homoscedastic variance performed much better in small samples. They are based on:

$$m_t^W(Y_t, X; \beta) = \frac{m_t(Y_t, X_t; \beta)}{\sqrt{E[m_t(Y_t, X_t; \beta)^2 | X = X_i]}} \,. \tag{7.17}$$

The expression of the conditional covariance matrix of these moments and the conditional expectation of the first derivatives are somewhat different from the previous ones, but the same general estimation principles can be applied in this case as well.[7] Inkman (2000) proposes additional Monte Carlo experiments comparing GMM estimators to SML with and without heteroskedasticity.

7.2.4.4 Other Estimators Based on Suboptimal Instruments

Of course there are many other specifications for the instrument matrix that lead to consistent, although not necessarily efficient, estimators for the slope coefficients. Their implementation as well as their efficiency ranking is discussed in detail in Bertschek and Lechner (1998). For example they show that the pooled probit estimator is asymptotically equivalent to the previous GMM estimator when the instruments are based on (7.13) to (7.16) but the off-diagonal elements of $\Omega(X_i)$ are set to zero. Avery, Hansen and Hotz (1983) also suggest to improve the efficiency of the pooled probit by exploiting strict exogeneity in another way by stacking the instrument matrix, so as to exploit that the conditional moment in period t is also uncorrelated with any function of regressors from other periods.

Chamberlain (1980) suggests yet another very simple route to improve the efficiency of the pooled probit estimator when there are arbitrary correlations of the errors over time which avoids setting up a 'complicated' GMM estimator. Since cross-section probits give consistent estimates of the coefficients for each period (scaled by the standard deviation of the period error term), the idea is to perform T probits period by period (leading to $T \times K$ coefficient estimates) and combine them in a second step using a Minimum Distance estimator. The variance–covariance matrix of estimators at different time periods should be computed to construct efficient estimates at the second step although small sample bias could also be a problem (Altonji Segal, 1996). In the case of homoscedasticity over time this step will be simple GLS, otherwise a nonlinear optimisation in the parametric space is required.[8]

[7] For all details, the reader is referred to Bertschek and Lechner (1998).

[8] Lechner (1995) proposes specification tests for this estimator.

7.2.5 How to Choose a Random Effects Estimator for an Application

This section introduced several estimators that are applicable in the case of random effect models under strict exogeneity. In practice the question is what correlation structure to impose and which estimator to use. Concerning the correlation structure, one has to bear in mind that exclusion restrictions are important for nonparametric identification and thus that explanatory variables should be sufficiently variable across time in order to permit the identification of a very general pattern of correlation of errors. For empirical applications of the estimators that we have reviewed, the following issues seem to be important: small sample performance, ease of computation, efficiency, robustness. We will address them in turn.

With respect to small sample performance of GMM estimators, Monte Carlo simulations by Breitung and Lechner (1997), Bertschek and Lechner (1998) and Inkmann (2000) suggest that estimators based on too many overidentifying restrictions (i.e. too many instruments), like the sequential estimators and some of the estimators suggested by Avery, Hansen and Hotz (1983) are subject to the typical weak instruments problem of GMM estimation due to too many instruments. Thus they are not very attractive for applications. The exactly identified estimators appear to work fine.

'Ease of computation' is partly a subjective judgement depending on computing skill and software available. Clearly, pooled probit is the easiest to implement, but random effects ML is available in many software packages as well. Exact ML is clearly not feasible for T larger than 4. For GMM and simulation methods, there is GAUSS code available on the Web (Geweke and Keane (2001) for instance) but they are not part of any commercial software package. The issue of computation time is less important now that it was some time ago (Greene, 2002) and the simulation estimators are getting more and more implementable with the increase of computing power. Asymptotic efficiency is important when samples are large. Clearly, exact ML is the most efficient one and can in principle be almost exactly approximated by the simulation estimators discussed.

With respect to robustness, it is probably most important to consider violations of the assumption that explanatory variables at all periods are exogeneous and restrictions of the autocorrelation structure of the error terms. We will address the issue of exogeneity at the end of this chapter though the general conclusions are very close to the linear case, as far as we know. Concerning the autocorrelation of errors, pooled probit either in its pseudo-ML or GMM version is robust if it uses marginal conditional moments. It is not true for the other ML estimators that rely on the correct specification of the autocorrelation structure. Finally, GMM estimators as they have been proposed here are robust against any autocorrelation. However, they obtain their efficiency gains by exploiting strict exogeneity and may become inconsistent if this assumption does not hold (with the exception of pooled probit, of course).

7.2.6 Correlated Effects

In the correlated effects (or unrelated effects) model, we abandon the assumption that individual effects and explanatory variables are independent. In analogy with the linear panel data case, Chamberlain (1984) proposes, in a random effect panel data nonlinear model, to replace the assumption that individual effects u_i are independent of the regressors by a weaker assumption. This assumption is derived from writing a linear regression:

$$u_i = X_i\gamma + \eta_i \tag{7.18}$$

where explanatory variables at all periods, X_i, are now independent of the redefined individual effect η_i. This parametrization is convenient but not totally consistent with the preceding assumptions: considering the individual effect as a function of the X_i variables makes its definition depend on the length of the panel. However, all results derived in the previous section can readily be applied by replacing explanatory variables X_{it} by the whole sequence X_i at each period.[9]

To recover the parameters of interest, β, two procedures can be used. The first method uses minimum distance estimation and the so called π-matrix technique of Chamberlain (Crépon and Mairesse, 1996). The reduced form:

$$y_{it}^{*} = X_i\gamma_t + \eta_i + v_{it}\ , \tag{7.19}$$

is first estimated. The second step consists in imposing the constraints given by:

$$\gamma_t = \gamma + \beta e_t \tag{7.20}$$

where e_t is an appropriate known matrix derived from (7.18) and (7.19).

The second procedure uses constrained maximum likelihood estimation by imposing the previous constraint (7.20) on the parameters of the structural model.

The assumption of independence between η_i and X_i is quite strong in the nonlinear case in stark contrast to the innocuous non-correlation assumption in the linear case. Moreover, it also introduces constraints on the data generating process of X_i if one wants to extend this framework when additional period information comes in Honoré (2002). Consider that we add a new period $T+1$ to the data and rewrite the projection as:

$$u_i = X_i\tilde{\gamma} + X_{iT+1}\tilde{\gamma}_{T+1} + \tilde{\eta}_i$$

By substracting both linear regressions at times T and $T+1$ and taking expectations conditional on information at period T, it implies that:

$$E(X_{iT+1} \mid X_i) = X_i(\gamma - \tilde{\gamma})/\tilde{\gamma}_{T+1}$$

which is not only linear in X_i but also only depend on parameters governing the y_{it} process.

[9] The so-called Mundlak (1978) approach is even more specific since individual effects u_i are written as a function of averages of covariates, $\frac{1}{T}\sum_{t=1}^{T} x_{it}$ only and a redefined individual effect η_i.

It is therefore tempting to relax (7.18) and admit that individual effects are a more general function of explanatory variables:

$$u_i = f(X_i) + \eta_i$$

where $f(.)$ is an unknown function satisfying weak restrictions (Newey, 1994). Even if the independence assumption between the individual effect η_i and explanatory variables X_i is still restrictive – because the variance of η_i is constant for instance – this framework is much more general than the previous one. What Newey (1994) proposes is based on the cross section estimation technique that we already talked about.

Consider the simple one-factor model where the variance of the individual-and-period specific shocks is not period-dependent, σ_v^2, and where the variance of η_i is such that $\sigma_v^2+\sigma_\eta^2$ is normalized to one. We therefore have:

$$E(y_{it} \mid X_i) = \Phi(X_{it}\beta + f(X_i))$$

where Φ is the distribution function of a zero-mean unit-variance normal variate. It translates into:

$$\Phi^{-1}(E(y_{it} \mid X_i)) = X_{it}\beta + f(X_i) \tag{7.21}$$

By any differencing operator (Arellano, 2003) and for instance by first differencing, we can eliminate the nuisance function $f(X_i)$ to get:

$$\Phi^{-1}(E(y_{it} \mid X_i)) - \Phi^{-1}(E(y_{it-1} \mid X_i)) = (X_{it} - X_{it-1})\beta \tag{7.22}$$

The estimation runs as follows. Estimates of $E(y_{it} \mid X_i)$ at any period are first obtained by series estimation (Newey, 1994) or any other nonparametric method (kernel, local linear, smoothing spline, see Pagan and Ullah, 1998 for instance). A consistent estimate of β is then obtained by using the previous moment condition (7.22).

A few remarks are in order. First, Newey (1994) proposes such a modeling framework in order to show how to derive asymptotic variance–covariance matrices of semi-parametric estimators. As it is outside of the scope of this chapter, the reader is refered, for this topic, to the original paper. It can also be noted that as an estimate of $f(X_i)$ can be obtained, in a second step, by using the equation in levels (7.21). One can then use a random effect approach to estimate the serial correlation of the random vector, v_{it}. Finally, there is a nonparametric version of this method (Chen, 1998) where Φ is replaced by an unknown function to be estimated, under some identification restrictions.

7.3 Fixed Effects Models Under Strict Exogeneity

In the so-called fixed effect model, the error component structure of Sect. 7.2.2 is assumed. The dependence between individual effects and explanatory variables is now unrestricted in contrast to the independence assumption in the random

effects model. In this section, we retain the assumption of strict exogeneity that explanatory variables and period-and-individual shocks are independent. We write the model as:

$$y_{it} = \mathbf{1}\{\mathbf{X}_{it}\beta + u_i + v_{it} > 0\} \tag{7.23}$$

where additional assumptions are developed below.

As the conditional distribution of individual effects u_i is unrestricted, the vector of individual effects should be treated as a nuisance parameter that we should either consistently estimate or that we should eliminate. If we cannot eliminate the fixed effects, asymptotics in T are required in most cases.[10] It is because only T observations are available to estimate each individual effect. It cannot be consistent as $N \to \infty$ and its inconsistency generically contaminates the estimation of the parameter of interest. It gives rise to the problem of incidental parameters (Lancaster, 2000). The assumption that T is fixed seems to be a reasonable approximation with survey data since the number of periods over which individuals are observed is often small. At the end of the section however, we will see how better large T approximations can be constructed for moderate values of T.

The other route is to difference out the individual effects. It is more difficult in non-linear models than in linear ones because it is not possible to consider linear transforms of the latent variable and to calculate within-type estimators. In other words, it is much harder to find moment conditions and specific likelihood functions that depend on the slope coefficient but do not depend on the fixed effects. In short panels, ML or GMM estimation of fixed effects probit models where the individual effects are treated as parameters to be estimated are severely biased if T is small (Heckman, 1981a).

In the first sub- sections we discuss some methods that appeared in the literature that circumvent this problem and lead to consistent estimators for $N \to \infty$ and T is *fixed*. Of course, there is always a price to pay either in terms of additional assumptions needed or in terms of the statistical properties of these estimators.

7.3.1 The Model

As already said, we consider (7.23) and we stick to the assumption of strict exogeneity of the explanatory variables:

$$F_{\varepsilon_t}(\varepsilon_{it}|u_i, X_{i1}, \ldots, X_{iT}) = F_{\varepsilon_t}(\varepsilon_{it}|u_i)\ . \tag{7.24}$$

Using the error component structure of Sect. 7.2.2, we can reformulate this assumption:

$$F_{v_t}(v_{it}|u_i, X_{i1}, \ldots, X_{iT}) = F_{v_t}(v_{it})\ . \tag{7.25}$$

[10] Not in all cases, the example of count data being prominent (Lancaster, 2000).

Note that $F_{\varepsilon_t}(\varepsilon_{it}|X_{i1},\ldots,X_{iT}) \neq F_{\varepsilon_t}(\varepsilon_{it})$ and also note that the distribution of the individual effect is unrestricted and can thus be correlated with observables. In most cases we will also impose that the errors are independent conditional on the fixed effect:

$$F(\varepsilon_{i1},\ldots,\varepsilon_{iT}|u_i,X_{i1},\ldots,X_{iT}) = \prod_{t=1}^{T} F_{\varepsilon_t}(\varepsilon_{it}|u_i) \tag{7.26}$$

$$F(v_{i1},\ldots,v_{iT}|u_i,X_{i1},\ldots,X_{iT}) = \prod_{t=1}^{T} F_{v_t}(v_{it}).$$

There are two obvious difficulties with respect to identification in such a model. First, it is impossible to identify the effects of time-invariant variables.[11] It has serious consequences because it implies that choice probabilities in the population are not identified. We cannot compare probabilities for different values of the explanatory variables. In other words, a fixed effect model that does not impose some assumption on distribution of the fixed effects cannot be used to identify causal (treatment) effects. This sometimes overlooked feature limits the use of fixed effects models.[12] What remains identified are the conditional treatment effects, conditional on any (unknown) value of the individual effect.

The second difficulty is specific to discrete data. In general, the individuals who stay all over the period of observation in a given state do not provide any information concerning the determination of the parameters. It stems from an identification problem, the so called mover-stayer problem. Consider someone which stays in state 1 from period 1 to T. Let v_i be any value of the individual-and-period shocks. Then if the individual effect u_i is a coherent value in model (7.23) with staying in the state all the time, then any value $\bar{u}_i \geq u_i$ is also coherent with model (7.23). Estimations are thus implemented on the sub-sample of people who move at least once between the two states ('moving' individuals).

7.3.2 The Method of Conditional Likelihood

The existence of biases leads to avoid direct ML estimations when the number of dates is less than ten (Heckman, 1981a). In certain cases, the bias can consist in multiplying by two the value of some parameters (Andersen, 1970; Chamberlain, 1984; Hsiao, 1996). This features makes this estimator pretty unattractive in large N, small T type of applications. If the logit specification is assumed however, it is possible to set up a conditional likelihood function whose maximisation gives consistent estimators of the parameters of interest β, regardless the length of the time period.

[11] It is however possible to define restrictions to identify these effects, see Chaps. 4 and 5.

[12] The claim that a parametric distributional assumption of individual effects is needed for the identification of causal treatment effects is however overly strong. What is true is that the estimation of the conditional distribution function of individual effects is almost never considered though it can be under much weaker assumptions than parametric ones.

Conditional logit: T periods

In the case where random errors, v_{it}, are independent over time and are logistically distributed, the sum $y_{i+} = \sum_{t=1}^{T} y_{it}$, is a sufficient statistic for the fixed effects, in the sense that the distribution of the data given y_{i+} does not depend on the fixed effect. Consider the logit model:

$$P(y_{it} = 1|X_i, u_i) = F(X_{it}\beta + u_i) \,, \tag{7.27}$$

where $F(z) = \frac{\exp(z)}{1+\exp(z)} = \frac{1}{1+\exp(-z)}$

The idea is to compute probabilities conditional on the number of times the individuals is in state 1:

$$L_i(\theta) = P\left(y_{i1} = \delta_{i1}, \ldots, y_{iT} = \delta_{iT} \mid X_i, u_i, \sum_{t=1}^{T} y_{it} = y_{i+}\right) = \frac{\exp\left(\sum_{t=1}^{T} X_{it}\beta\,\delta_{it}\right)}{\sum_{d \in B_i} \exp\left(\sum_{t=1}^{T} X_{it}\beta d_t\right)}$$

where

$$B_i = \left\{ d = (d_1, \ldots, d_T) \text{ such that } d_t \in \{0,1\} \text{and} \sum_{t=1}^{T} d_t = \sum_{t=1}^{T} y_{it} \right\}$$

The set B_i differs between individuals according to the value of $\sum_{t=1}^{T} y_{it}$, i.e., the number of visits to state 1. Parameter β is estimated by maximising this conditional log-likelihood function. The estimator is consistent as $N \to \infty$, regardless of T (Andersen, 1970, Chamberlain, 1980, 1984, Hsiao, 1996). Nothing is known about its efficiency as in general conditional likelihood estimators are not efficient. Note that only the 'moving' individuals are used in the computation of the conditional likelihood. Extensions of model (7.27) can be considered. For instance, Thomas (2003) develops the case where individual effect are multiplied by a time effect which is to be estimated.

The estimation of such a $T-$period model is also possible by reducing sequences of T observations into pairs of binary variables. Lee (2002) develop two interesting cases. First, the T periods can be chained sequentially two-by-two and a $T = 2$ conditional model can be estimated (as in Manski, 1987 see below). All pairs of periods two-by-two could also be considered. These decompositions will have an interest when generalizing conditional logit, when considering semi-parametric methods or more casually, as initial conditions for conditional maximum likelihood. It is why we now review the $T = 2$ case.

7.3.2.1 An Example: The Two Period Static Logit Model

The conditional log-likelihood based on the logit model with T=2 computed with *moving* individuals is given by:

$$L = \sum_{d_i=1} \log \frac{\exp(X_{i2}\beta)}{\exp(X_{i1}\beta)+\exp(X_{i2}\beta)} + \sum_{d_i=0} \log \frac{\exp(X_{i1}\beta)}{\exp(X_{i1}\beta)+\exp(X_{i2}\beta)} ,$$

where for *moving* individuals, the binary variable d_i is:

$$\begin{cases} d_i = 1 \text{ if } y_{i1} = 0, y_{i2} = 1 \\ d_i = 0 \text{ if } y_{i1} = 1, y_{i2} = 0 \end{cases}$$

Denote $\Delta X_i = X_{i2} - X_{i1}$. The conditional log-likelihood becomes:

$$L = \sum_{i|d_i=1} \log \frac{\exp(\Delta X_i\beta)}{1+\exp(\Delta X_i\beta)} + \sum_{i|d_i=0} \log \frac{1}{1+\exp(\Delta X_i\beta)}$$

which is the expression of the log-likelihood of the usual logit model:

$$P(d_i = 1|\Delta X_i) = F(\Delta X_i\beta) \tag{7.28}$$

adjusted on the sub-sample of *moving* individuals. Note that the regressors do not include an intercept, since in the original model the intercept was absorbed by the individual effects.

7.3.2.2 A Generalization

The consistency properties of conditional likelihood estimators are well known (Andersen, 1970) and lead to the interesting properties of conditional logit. This method has however been criticized on the ground that assuming a logistic function is a strong distributional assumption. When the errors v_{i1} and v_{i2} are independent, it can be shown that the conditional likelihood method is applicable only when the errors are logistic (Magnac (2004)). It is possible however to relax the independence assumption between errors v_{i1} and v_{i2} to develop a richer semi-parametric or parametric framework in the case of two periods. As above, pairing observations two-by-two presented by Lee (2002) can be used when the number of periods is larger.

The idea relies on writing the condition that the sum $y_{i1} + y_{i2} = 1$ is a sufficient statistic in the sense that the following conditional probability does not depend on individual effects:

$$P\left(y_{i1} = 1, y_{i2} = 0 \mid X_i, u_i, \sum_{t=1}^{2} y_{it} = 1\right) = P\left(y_{i1} = 1, y_{i2} = 0 \mid X_i, \sum_{t=1}^{2} y_{it} = 1\right)$$

In that case, the development in the previous section can be repeated because the conditional likelihood function depends on parameter β and not on individual effects. It can be shown that we end up with an analog of (7.28) where distribution $F(.)$ is a general function which features and semi-parametric estimation are discussed in Magnac (2004).

7.3.3 Fixed Effects Maximum Score

The methods discussed until Sect. 3.2.2 are very attractive under one key condition, namely that the chosen distributional assumptions for the latent model are correct, otherwise the estimators will be typically inconsistent for the parameters of the model. However, since those functional restrictions are usually chosen for computational convenience instead of a priori plausibility, models that require less stringent assumptions or which are robust to violations of these assumptions, are attractive. Manski (1987) was the first to suggest a consistent estimator for fixed effects models in situations where the other approaches do not work. His estimator is a direct extension of the maximum score estimator for the binary model (Manski, 1975). The idea of this estimator for cross-sectional data is that if the *median* of the error term conditional on the regressors is zero, then observations with $X_i\beta > 0$ (resp. < 0) will have $P(y = 1\,|X_i\beta > 0) > 0.5$ (resp < 0.5). Under some regularity conditions this implies that $E\{\text{sgn}(2y_i - 1)\text{sgn}(X_i\beta)\}$ is uniquely maximised at the true value (in other words $(2y_i - 1)$ and $(X_i\beta)$ should have the same sign). Therefore, the analogue estimator obtained by substituting expectations by means is consistent although not asymptotically normal and converges at a rate $N^{1/3}$ to a non-normal distribution (Kim and Pollard, 1990). There is however a smoothed version of this estimator where the sign function is substituted with a kernel type function, which is asymptotically normal and comes arbitrarily close to $\sqrt{N}$-convergence if tuning parameters are suitably chosen (Horowitz, 1992). However, Chamberlain (1992) shows that it is not possible of attaining a rate of $\sqrt{N}$ in the framework adopted by these papers.

Using a similar reasoning as in the conditional logit model and using the assumption that the distribution of the errors over time is stationary, Manski (1987) showed that, conditional on X:

$$P(y_2 = 1\,|y_2 + y_1 = 1, X) > 0.5 \text{ if } (X_2 - X_1)\beta > 0$$

Therefore, for a given individual higher values of $X_t\beta$ are more likely to be associated with $y_t = 1$. In a similar fashion as the cross-sectional maximum score estimator, this suggests the following conditional maximum score estimator:

$$\hat{\beta}_N = \arg\max_{\beta} \sum_{i=1}^{N} \text{sgn}(y_{i2} - y_{i1})\text{sgn}[(X_{i2} - X_{i1})\beta]$$

For longer panels one can consider all possible pairs of observations over time:

$$\hat{\beta}_N = \arg\max_{\beta} \sum_{i=1}^{N} \sum_{s<t} \text{sgn}(y_{is} - y_{it}) \text{sgn}[(X_{is} - X_{it})\beta]$$

The estimator has similar properties than the cross-sectional M-score estimator, in the sense that it is consistent under very weak conditions, but not asymptotically normal and converges at a rate slower than $\sqrt{N}$. Kyriazidou (1995) and Charlier, Melenberg and van Soest (1995) show that the same 'smoothing trick' that worked for the cross-sectional M-score estimator also works for the conditional panel version. Hence, depending on the choice of smoothing parameters, the rate of convergence may come arbitrarily close to $\sqrt{N}$.

In practice, there are few applications of this estimator, since many difficulties arise: the solution of the optimisation problem is not unique, and the optimisation can be very complicated, because of the step function involved.

Other semi-parametric methods of estimation include Lee (1999) and Honoré and Lewbel 2002). In the first paper, an assumption about the dependence between individual effects and explanatory variables allows for the construction of the method of moments estimator which is root-N consistent and asymptotically normal. In the second paper, another partial independence assumption is made as well as assumptions about the large support of one special continuous covariate. By linearizing the model (Lewbel, 2000), one can return to the reassuring world of linear models and difference out the individual effects. The reader is referred to the original papers in both cases.

7.3.4 GMM Estimation

A possible solution to solving the problem posed by the presence of unobservable individual effect is to propose moment conditions which will be approximately satisfied provided that the individual effects are small, and estimators based on such moments (Laisney and Lechner, 2002). Consider the moment condition for any $t = 1, \ldots, T$:

$$E(y_t \,|X_i, u_i) = F(x_{it}\beta + u_i)$$

When the individual effect is close enough to the value of $\tilde{u}$, the first order Taylor approximation around $u = \tilde{u}$ is exact, so we can write for any $s,t = 1, \ldots, T$:

$$U - \tilde{u} = \frac{E[y_t \,|X, u] - F(X_t\beta + \tilde{u})}{f(X_t\beta + \tilde{u})} = \frac{E[y_s \,|X, u] - F(X_s\beta + \tilde{u})}{f(X_s\beta + \tilde{u})}$$

Thus, for any $s,t = 1, \ldots, T; s \neq t$, the following function,

$$m_{ts}(y, X; \beta) = \frac{y_t - F(X_t\beta - \tilde{u})}{f(X_t\beta - \tilde{u})} - \frac{y_s - F(X_s\beta - \tilde{u})}{f(X_s\beta - \tilde{u})}$$

has a conditional mean of zero at the true value of β, given $X = X_i$. It can be used as the basis for (almost) consistent estimation of the panel probit model with fixed effects close to $\tilde{u}$. Under standard regularity conditions, a GMM estimator of the coefficients for the time varying regressors of the panel model based on these moment functions is consistent (almost, given the Taylor approximation) and $\sqrt{N}$ asymptotically normal (Newey, 1993; Newey and McFadden, 1994).

7.3.5 Large-*T Approximations*

Finally, there are some new developments that are only briefly sketched here and that rely on large-T approximations in parametric binary models. The inspiration comes from Heckman (1981a) pioneering work. Monte Carlo experiments can indeed be used to assess the magnitude of the bias of fixed effect estimators in binary probit or logit models as it was developed in the previous section. This bias due to the presence of incidental parameters is of order $O(T^{-1})$ in panel probit and for values around $T = 10$ the bias is found to be small (see also Greene, 2002).

A first direction for improving estimators is to assess and compute the bias either analytically or by using jackknife techniques as proposed by Hahn and Newey (2004). Under assumptions of independence over time of regressors and disturbances, bias-corrected estimators can be easily constructed. Hahn and Kuesteiner (2004) relax the assumption of independence over time by proposing another analytical correction of the bias and that could also apply to the dynamic case (see next section).

The second direction relies on parameter orthogonalization. Inconsistency of fixed effects estimators occurs because the number of useful observations to estimate individual effects is fixed and equal to T and because there is contamination from the inconsistency of individual effect estimates into the parameters of interest. If, as in the Poisson count data example,[13] parameters of interest and individual effects can be factored out in the likelihood function (Lancaster, 2003) contamination is absent. Parameters are said to be orthogonal. These cases are not frequent however. The pionnering work of Cox and Reid (1987) uses a weaker notion of information orthogonality. At the true parameter values, the expectation of the cross derivative of the likelihood function w.r.t. the parameter of interest and the nuisance parameters is equal to zero. The invariance of likelihood methods to reparametrizations can then be used. The reparametrization which is interesting to use is the one (if it exists) that leads to information orthogonality. If this reparametrization is performed and if the nuisance parameters are integrated out in Bayesian settings, or concentrated out in classical settings, the bias of the ML estimator is of order $1/T^2$ instead of $1/T$ (in probability). For Probit (or other parametric) models, this method is proposed by Lancaster (2003) in a Bayesian setting. General theory in parametric non linear models in the Bayesian case is developed by Woutersen (2002). In the classical case, the panel static probit model is studied in a Monte Carlo experiment

[13] As described in Montalvo (1997) and Blundell, Griffith and Windmeijer (2002).

as an example by Arellano (2003) and in a dynamic case by Carro (2003). They show that for moderate T (4.6), the bias is small. It is smaller than the value for T advocated by Heckman (1981a) though these values shall be theoretically validated in each instance where it is applied, as always when using Monte Carlo experiments about approximations.

7.4 Dynamic Models

In dynamic models where explanatory variables comprise lagged endogenous variables and other predetermined variables, we could further abandon the assumption that individual-and-period shocks and explanatory variables are independent. We distinguish again random and fixed effects models. This section is short not because the subject is unimportant but because the main ideas are extensions of the strict exogeneity case. There is one original issue however that we shall insist on, which is the choice of initial conditions.

7.4.1 Dynamic Random Effects Models

There are many potential sources of dynamics in econometric models. Some sources are easily dealt with in the framework of the last section: coefficients changing over time, lagged values of the strictly exogenous explanatory variables, correlation of random effects over time. There could also be true state dependence that is structural dependence on the lagged dependent variable or feedback effects of dependent variables on explanatory variables. Those explanatory variables are thus predetermined instead of strictly exogeneous. Most behavioral economic models using time-series or panel data are likely to be dynamic in this sense.

There are various dynamic discrete models as introduced by Heckman (1981a). The latent model that we study in this section, is written as:

$$y_{it}^{*} = \alpha y_{it-1} + X_{it}\beta + u_i + v_{it} \tag{7.29}$$

where individual effects u_i or individual-and-period specific effects v_{it} are or can be dependent of explanatory variables y_{it-1} and X_{it} and/or the future of these variables. It is in this sense that right-hand side variables are endogenous in this section. For simplicity we here consider one lag only and that v_{it} are independent of the past and present of (y_{it-1}, X_{it}).

As an alternative to this model (7.29), there is a class of models in which the lagged latent variable, y_{it-1}^{*}, is included among explanatory variables instead of the binary variable y_{it}. This type of dynamics is called habit persistence. Because recursive substitution techniques can be used – the lagged latent variable is replaced recursively by their expression (7.29) – these habit persistence models can

be transformed into static models where explanatory variables include lags of the exogenous variables and where some care should be taken with the initial condition, y_{i1}^*. These types of models are discussed briefly in Heckman (1981b). Estimation of the structural parameters in the case of binary choice is detailed in Lechner (1993). Moreover, this framework does not accomodate weak endogeneity which is one of the focus of this section.

7.4.1.1 Initial Conditions

When the lagged endogenous variable is present, there is an initial condition problem as in the linear case though it is more diffcult to deal with. Assuming for the moment that there are no other explanatory variables, $\beta = 0$, the likelihood function is written by conditioning on individual effects as in the previous section:

$$l(y_{iT}, ., y_{i2}, y_{i1} \mid u_i) = \prod_{t=2}^{T} l(y_{it} \mid y_{it-1}, u_i) l(y_{i1} \mid u_i)$$

It is obvious that one needs additional information for deriving $l(y_{i1} \mid u_i)$ that model (7.29) is not providing. It is analogous to the linear case and the assumptions that initial conditions are exogenous or that initial conditions are obtained by initializing the process in the infinite past were soon seen to be too strong or misplaced. They are generally strongly rejected by the data. Heckman (1981) proposed to use an auxiliary assumption such as:

$$y_{i1}^* = \theta u_i + v_{i1}^0 \ . \tag{7.30}$$

The complete likelihood function is then obtained by integrating out, u_i, as before.

Another route was suggested by Wooldridge (2002) or Arellano and Carrasco (2003). Instead of using the complete joint likelihood function, they resort to the following conditional likelihood function:

$$l(y_{iT}, ., y_{i2} \mid y_{i1}, u_i) = \prod_{t=2}^{T} l(y_{it} \mid y_{it-1}, u_i) \ .$$

When integrating out u_i, one now needs to choose the conditional distribution function $f(u_i \mid y_1)$ which might be written as the auxiliary model which marries well with the approach of Chamberlain seen above:

$$u_i = \theta y_{i1} + \eta_i \tag{7.31}$$

It should be noted that one loses information and that it is not immediately clear whether restriction (7.30) is more restrictive than (7.31) in particular when other explanatory variables are present in the model.

7.4.1.2 Monte Carlo Experiments of Simulation Methods

In the literature, some papers report Monte Carlo experiments of random effects dynamic models estimated by simulation (Keane, 1994; Chib and Jeliazkov, 2002; Lee, 1997). There seems to be a consensus on a few results. Estimates of the autoregressive parameter seem to be downward biased while parameters of the variance of random effects can be upward or downward biased according to the model (Lee, 1997). Biases increase when serial correlation is stronger though it can be counteracted by increasing the number of draws either for SML or MSM as well as for Gibbs sampling. Biases also increase when the number of periods increases. Misspecification of initial conditions introduces fairly large biases in the estimation.

7.4.1.3 A Projection Method

For treating the weakly endogenous case, there has been an interesting suggestion proposed by Arellano and Carrasco (2003). Let $\omega_{it} = (y_{it-1}, X_{it})$ be the relevant conditional information in period t that is grouped into the information set, $\omega_i^t = (\omega_{it}, \omega_i^{t-1})$ where ω_i^0 is the empty set. Variables ω_i^t summarize the relevant past of the process until period t, that is the sequence of lagged endogeneous variables, explanatory variables and their lags and any other piece of information such as instruments for instance. Assume that $\varepsilon_{it} = u_i + v_{it}$ is such that:

$$\varepsilon_{it} \mid \omega_i^t \rightsquigarrow N(E(u_i \mid \omega_i^t), \sigma_t^2)$$

where independence between v_{it} and the information set ω_i^t has been used. Thus, it rules out serial correlation in the usual sense[14] while allowing for feedback. It thus constitutes a generalization of the setting of the projection method of Chamberlain (1980) and Newey (1994) that we presented in the previous section.

The sequence of conditional means $E(u_i \mid \omega_i^t)$ are related by the moment conditions:

$$E(E(u_i \mid \omega_i^t) \mid \omega_i^{t-1}) = E(u_i \mid \omega_i^{t-1}) \tag{7.32}$$

Write now the conditional means:

$$E(y_{it} \mid \omega_i^t) = \Phi\left(\frac{\alpha y_{it-1} + X_{it}\beta + E(u_i \mid \omega_i^t)}{\sigma_t}\right)$$

which translates into:

$$\sigma_t . \Phi^{-1}(E(y_{it} \mid \omega_i^t)) = \alpha y_{it-1} + X_{it}\beta + E(u_i \mid \omega_i^t)$$

[14] Individual-and-period v_{it-1} is not included in ω_i^t, only y_{it-1} is.

The moment condition (7.32) is thus:

$$E(\sigma_t.\Phi^{-1}(E(y_{it} \mid \omega_i^t)) - (\alpha y_{it-1} + X_{it}\beta) \mid \omega_i^{t-1}) = \\ = \sigma_{t-1}.\Phi^{-1}(E(y_{it} \mid \omega_i^{t-1})) - (\alpha y_{it-2} + X_{it-1}\beta)$$

As before, some nonparametric estimates of $E(y_{it} \mid \omega_i^t)$ can be obtained and plugged in this moment condition.

As it is formally identifical to the approach proposed by Newey (1994), the same remarks can be addressed to this approach. There may however be a curse of dimensionality coming in because the dimension of ω_i^t is growing with the number of periods. Arellano and Carrasco (2003) proposes simplifications and the reader is referred to the original paper.

7.4.2 Dynamic Fixed Effects Models

Chamberlain (1985) extends the conditional logit method to the case where the lagged endogenous variable is the only covariate (see also Magnac, 2000, for multinomial and dynamic models where lags can be larger than 1). Sufficient statistics are now a vector of three variables. On top of the sum of binary variables, the binary variables at the first and last period are added to the list. For instance, in the case where only one lag is used, the smallest number of periods for identification is equal to 4 and the useful information is contained in the intermediate periods from $t = 2$ to $T - 1$. The main drawback of this method is that, in the logit case and in the model with one lag, the sum of binary variables, the first and last values of the binary variables are not sufficient statistics if other explanatory variables are present in the model.

If explanatory variables are discrete, the idea proposed by Honoré and Kyriazidou (2000) is to consider only the observations such that explanatory variables are constant in the intermediate periods from $t = 2$ to $T - 1$. Conditional to the values of these explanatory variables, the sum of binary variables, the first and last values of the binary variables are now sufficient statistics. In order to accomodate continuous variables, Honoré and Kyriazidou (2000) proposes to use observations such that explanatory variables are approximately constant in the intermediate periods from 2 to $T - 1$. The statistics described above are approximately sufficient. Observations can be weighted according to the degree of such an approximation. Under some conditions the estimator is consistent and asymptotically normal, but due to the nonparametric part, its convergence rate is less than $\sqrt{N}$. Note also that this construction rules out time dummies, which cannot by definition be similar in two periods.

References

Altonji, J.G., and L.M., Segal, 1996, "Small-sample bias in GMM estimation of covariance structures", *Journal of Business Economics and Statistics*, 14: 353–366.

Andersen, E.B., 1970, "Asymptotic properties of conditional maximum likelihood estimators", *Journal of the Royal Statistic Society*, Series B, 32: 283–301.

Angrist, J.D., 2001, "Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice", *Journal of Business Economics and Statistics,* 19:2–16.

Arellano, M., 2003, "Discrete choice with panel data", *Investigaciones Economicas,* 27, 423–458.

Arellano, M., and R., Carrasco, 2003, "Binary choice panel data Models with Predetermined Variables", *Journal of Econometrics*, 115, 125–157.

Arellano, M., and B., Honoré, 2001, "Panel data models: Some recent developments", in J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics,* North Holland: Amsterdam V(53):3229–3296.

Avery, R.B, L.P., Hansen, and V.J., Hotz, 1983, "Multiperiod probit models and orthogonality condition estimation", *International Economic Review*, 24:21–35.

Baltagi, B.H., 2000, *Econometric Analysis of Panel Data,* Wiley: London.

Butler, J., and R., Moffitt, 1982, "A computationally efficient quadrature procedure for the one-factor multinomial probit model ", *Econometrica*, 50(3): 761–764.

Bertschek, I., and M., Lechner, 1998, "Convenient estimators for the panel probit model", *Journal of Econometrics*, 87: 329–371.

Blundell, R., R., Griffith, and F., Windmeijer, 2002, "Individual Effects and Dynamics in Count data Models", *Journal of Econometrics,* 108: 113–131.

Breitung, J., and M., Lechner, 1997, "Some GMM estimation methods and specification tests for nonlinear models", in L. Mátyás and P.Sevestre (Eds.), *The Econometrics of Panel Data*, 2nd ed., Dordrecht: Kluwer, 583–612, 1996.

Carro, J.M., 2003, "Estimating dynamic panel data discrete choice models with fixed effects", Working paper, CEMFI, 0304.

Chamberlain, G., 1980, "Analysis of covariance with qualitative data", *Review of Economic Studies*, 47: 225–238.

Chamberlain, G., 1984, "Panel Data", in Z. Griliches and M.D. Intrilligator (Eds.), *Handbook of Econometrics*, vol II, ch 22, Elsevier Science: Amsterdam, 1248–1318.

Chamberlain, G., 1985, "Heterogeneity, omitted variable bias and duration dependence", in *Longitudinal Analysis of Labor Market Data,* in J.J. Heckman and B. Singer (Eds.), Cambridge UP: Cambridge.

Chamberlain, G., 1992, "Binary response models for panel data: Identification and information", *Mimeo*, Harvard University: Cambridge.

Charlier, E., B., Melenberg, and A., van Soest, 1995, "A smoothed maximum score estimator for the binary choice panel model and an application to labour force participation", *Statistica Neerlandica*, 49: 324–342.

Chen, S., 1998, "Root-N consistent estimation of a panel data sample selection model", unpublished manuscript, Hong Kong university.

Chib, S., 2001, "Markov Chain Monte Carlo Methods: Computation and Inference", in J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics,* V(57):3570–3649.

Chib, S., and E., Greenberg, 1998, "Analysis of multivariate probit models", *Biometrika,* 85:347–61.

Chib, S., and I., Jeliazkov, 2002, "Semiparametric hierarchical bayes analysis of discrete panel data with state dependence", Washington University, working paper.

Cox, D.R., and M., Reid, 1987, "Parameter orthogonality and approximate conditional inference", *Journal of the Royal Statistical Society*, Series B, 49:1–39.

Crépon, B., and J., Mairesse, 1996, "The chamberlain approach to panel data: An overview and some simulation experiments", in L. Matyas and P. Sevestre (Eds.), *The Econometrics of Panel Data*, Kluwer: Amsterdam.

Geweke, J., M., Keane, and D.E., Runkle, 1997, "Statistical inference in the multinomial multiperiod probit model", *Journal of Econometrics*, 80, 125–165.

Geweke, J.F., and M., Keane, 2001, "Computationally intensive methods for integration in econometrics", in J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics,* V(56):3465–3568.

Gouriéroux, C., and A., Monfort, 1993, "Simulation-based inference: A survey with special reference to panel data models", *Journal of Econometrics*, 59: 5–33.

Gouriéroux, C., and A., Monfort, 1996, *Simulation-based Econometric Methods*, Louvain: CORE Lecture Series.

Gouriéroux, C., A., Monfort, and A., Trognon, 1984, "Pseudo-likelihood methods - Theory", *Econometrica,* 52: 681–700.

Greene, W., 2002, "The Bias of the fixed effects estimator in non linear models", New York University: New York, unpublished manuscript.

Greene, W., 2003, *Econometric Analysis,* 5th ed., Prentice Hall: Englewood Cliffs.

Guilkey, D.K., and Murphy, J.L., 1993, "Estimation and testing in the random effects probit model", *Journal of Econometrics*, 59: 301–317.

Hahn, J., and G., Kuersteiner, 2004, "Bias reduction for dynamic nonlinear panel models with fixed effects", MIT unpublished manuscript.

Hahn, J., and W., Newey, 2004, "Jackknife and analytical Bias reduction for nonlinear panel data models", *Econometrica,* 72:1295–1319.

Hajivassiliou, V., and D., McFadden, 1998, "The method of simulated scores for the estimation of LDV models ", *Econometrica*, 66: 863–896.

Hajivassiliou, V., D., McFadden, and P., Ruud, 1996, "Simulation of multivariate normal rectangle probabilities and their derivatives. Theorical and computational results", *Journal of Econometrics*, 72: 85–134.

Heckman, J.J., 1981a, "The incidental parameters problem and the problem of initial conditions in estimating a discrete time – discrete data stochastic process and some Monte-Carlo evidence," in C. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data*, MIT Press, Cambridge, MA, 179–195.

Heckman, J.J., 1981b, "Statistical models for discrete panel Data" in C. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data*, MIT Press, Cambridge, MA, 114:178.

Heckman, J.J., and B., Singer, 1984, "A method for minimizing the impact of distributional assumptions in econometric models for duration data", *Econometrica*, 52:271–320.

Honoré, B., 2002, "Non-linear models with panel data", WP CEMMAP, 13/02.

Honoré, B., and E., Kyriazidou, 2000, "Panel data discrete choice modles with lagged dependent variables", *Econometrica,* 68:839–874.

Honoré, B.E., and A., Lewbel, 2002, "Semiparametric binary choice panel data models without strict exogeneity", *Econometrica*, 70:2053–2063.

Horowitz, J., 1992, "A smoothed maximum score estimator for the binary response model", *Econometrica*, 60: 505–531.

Hsiao, C., 1992, "Logit and Probit Models", in L. Mátyás and P. Sevestre (Eds.), *The Econometrics of Panel Data: Handbook of Theory and Applications*, Chap. 11: 223–241, Kluwer: Amsterdam.

Hsiao, C., 1996, "Logit and probit models", in L. Mátyás and P. Sevestre (Eds.), *The Econometrics of Panel Data: Handbook of Theory and Applications*, 2nd ed., Chap. 16: 410–428, Kluwer: Amsterdam.

Hsiao, C., 2003, Analysis of panal data, 2nd ed., Cambridge University Press, Econometric Society Monographs, 11.

Inkman, J., 2000, "Misspecified heteroskedasticity in the panel probit model: A small sample comparison of GMM and SML estimators", *Journal of Econometrics*, 97: 227–259.

Kamionka, T., 1998, "Simulated maximum likelihood estimation in transition models", *Econometrics Journal,* 1:C129–153.

Keane, M.P., 1994, "A computationally efficient practical simulation estimator for panel data", *Econometrica*, 62:95–116.

Kim, J., and D., Pollard, 1990, "Cube root asymptotics", *Annals of Statistics*, 18: 191–219.

Kyriazidou, E., 1995, *Essays in Estimation and Testing of Econometric Models*, Ph.D. dissertation, Northwestern University.

Laisney, F., and M., Lechner, 2002, "Almost consistent estimation of panel probit models with 'Small' fixed effects", *Discussion paper* no. 2002–15, University of St. Gallen.

Lancaster, A., 2000, "The incidental parameter problem since 1948", *Journal of Econometrics,* 95:391–413.

Lancaster, A., 2003, *An Introduction to Modern Bayesian Econometrics,* Blackwell: Oxford.

Lechner, M., 1993, "Estimation of limited dependent variable habit persistence models on panel data with an application to the dynamics of self-employment in the former east germany", in H. Bunzel, P. Jensen, and N. Westergård-Nielson, (Eds.), *Panel Data and Labour Market Dynamics*,: North-Holland Amsterdam, 263–283.

Lechner, M., 1995, "Some specification tests for probit models estimated on panel data", *Journal of Business & Economic Statistics*, 13: 475–488, 1995.

Lee, L.F., 1992, "On efficiency of methods of simulated moments and maximum simulated likelihood estimation of discrete response models", *Econometric Theory*, 8:518–552.

Lee, L.F., 1995, "Asymptotic bias in simulated maximum likelihood estimation of discrete choice models", *Econometric Theory*, 11:437–483.

Lee, L.F., 1997, "Simulated maximum likelihood estimation of dynamic discrete choice statistical models: Some Monte carlo results", *Journal of Econometrics*, 82:1–35.

Lee, L.F., 2000, "A numerically stable quadrature procedure for the one-factor random component discrete choice model", *Journal of Econometrics*, 95: 117–129.

Lee, M.J., 1999, "A root-n consistent semiparametric estimator for related-effect binary response panel data", *Econometrica*, 67:427–33.

Lee, M.J., 2002, *Panel Data Econometrics,* Academic Press: New York.

Lewbel, A., 2000, "Semiparametric qualitative response model estimation with unknown Heteroskedasticity or instrumental variables", *Journal of Econometrics*, 97:145–177.

McFadden, D., 1989, "A method of simulated moments for estimation od discrete response models without numerical integration", *Econometrica*, 57: 995–1026.

Magnac, T., 2000, "State dependence and unobserved heterogeneity in youth employment histories", *The Economic Journal,* 110:805–837.

Magnac, T. 2004, "Binary Variables and Sufficiency: Generalizing the Conditional Logit", *Econometrica,* 72:1859–1876.

Manski, C.F. 1975, "Maximum Score Estimation of the Stochastic Utility Model", *Journal of Econometrics*, 3, 205–228.

Manski, C.F., 1987, "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data", *Econometrica*, 55, 357–362.

Montalvo, J.G., 1997, "GMM estimation of count-panel-data models with Fixed Effects and Predetermined Instruments", *Journal of Business Economics and Statistics*, 15: 82–89.

Mroz, T., 1999, "Discrete factor approximation in simultaneous equation models: Estimating the impact of a dummy endogenous variable on continuous outcome", *Journal of Econometrics*,92: 233–274.

Mundlak, Y., 1978, "On the Pooling of Time Series and Cross Section Data", *Econometrica*, 46: 69–85.

Newey, W., 1993, "Efficient estimation of models with conditional moment restrictions", in G.S. Maddala, C. Rao, Vinod, H. (Eds.), *Handbook of Statistics*, Vol. 11, Ch. 16, North-Holland: Amsterdam.

Newey, W., 1994, "The asymptotic variance of semiparametric estimators", *Econometrica*, 62:1349–1382.

Newey, W.K., and McFadden, D., 1994, "Large sample estimation and hypothesis testing", in R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics*, Vol. 4, 2113–2245, North-Holland: Amsterdam.

Pagan, A., and A., Ullah, 1998, *Nonparametric Econometrics,* Cambridge UP, Cambridge.
Robinson, P.M., 1982, "On the asymptotic properties of estimators of models containing limited dependent variables", *Econometrica*, 50:27–41.
Sevestre, P., 2002, *Econométrie des données de panel*, Dunod: Paris.
Thomas, A., 2003, "Consistent estimation of binary-choice panel data models with heterogeneous linear trends", LEERNA-INRA Toulouse, unpublished manuscript.
Train, K., 2002, *Discrete Choices with Simulation,* Cambridge UP:Cambridge.
Wooldridge, J., 2000, *Introductory Econometrics*, 2nd ed., South-Western College Publishing: Boston, MA.
Wooldridge, J., 2002, "Simple solutions to the initial conditions problem in dynamic non linear panel data models with unobserved heterogeneity", WP CEMMAP, London, 18/02.
Woutersen T., 2002, "Robustness against incidental parameters", Western Ontario, unpublished manuscript.

Part II
Advanced Topics

Chapter 8
Dynamic Models for Short Panels

Mark N. Harris, László Mátyás and Patrick Sevestre

8.1 Introduction

The empirical analysis of economic behavior often entails specifying dynamic econometric models; that is, models with lagged dependent variable(s) among the regressors. As in time-series context, when the model is dynamic, standard estimation methods based on least squares generally do not lead to estimators having good properties. Indeed, for dynamic panel data models, methods such as OLS or the Within estimators are not consistent. This results from the fact that, due to the unobserved effects, the lagged dependent variable and the disturbance terms are correlated.

Therefore one has to resort to alternative methods. The most commonly used approach is that of GMM, relying on a properly defined set of instrumental variables, or equivalently, a set of orthogonality conditions (see Chap. 4). A large part of this chapter is devoted to the presentation of a number of such consistent estimators. We consider different sets of assumptions about the exogeneity of the regressors (other than the lagged dependent variable), focusing in particular on their possible correlation with the individual effects. We also consider maximum likelihood estimation, although it is less frequently used in practice as its computation is typically much more complex. Finally, besides the estimation of unknown parameters, hypothesis testing is also considered.

Mark N. Harris
Department of Econometrics & Business Statistics, Monash University, Victoria 3800, Australia, e-mail: mark.harris@buseco.monash.edu.au

László Mátyás
Central European University, Department of Economics, Nádor u. 9, 1051 Budapest, Hungary, e-mail: matyas@ceu.hu

Patrick Sevestre
Université Paris 1 – Panthéon Sorbonne, Ecole Economique de Paris (Paris School of Economics), 106-112 Boulevard de l'Hôpital, 75013 Paris, France, e-mail: sevestre@univ-paris1.fr

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

It is worth noting that, given the typical dimensions of most microeconomic panel data sets, the focus of this chapter is on "short panels", i.e., panels with a large number of observations on economic units (N) over a limited number of time periods (T).[1] Moreover, some guidelines for practitioners are also given about the finite sample behavior of the proposed methods by summing up some available Monte Carlo simulation evidence as well as a few other finite sample results (e.g., Kiviet, 1995).

8.2 The Model

Consider that the economic behavior of interest can be represented by the following regression model

$$E(y|y_{-1}, X, \alpha) = \delta y_{-1} + X\beta + \alpha, \tag{8.1}$$

where y is the variable of interest. It is determined by: its past realisation, y_{-1}; a matrix of observed characteristics, X (typically including a constant term); and an unobserved, or "individual", effect, α.[2] δ (a scalar) and β (a vector) are the unknown coefficients to be estimated. Although X may contain both time-variant and time-invariant observed characteristics, it will be implicitly assumed throughout this chapter that it contains only the former.[3] Note also, for ease of exposition, we assume that the panel is balanced (each economic unit is observed over the same time period).[4] Due to the dimensions of typical panel data sets, we assume a random effects approach throughout. However, we do not rule out the possibility of a correlation between those random individual effects and the regressors, a situation which is quite close to that associated with a fixed effects approach.

The sample regression function associated with the above model can be written as

$$y_{it} = \delta y_{i,t-1} + x'_{it}\beta + \alpha_i + \varepsilon_{it},\ i = 1, \ldots, N;\ t = 1, \ldots, T, \tag{8.2}$$

where the α_i's are i.i.d. $(0, \sigma_\alpha^2)$ and the ε_{it}'s are idiosyncratic error terms, also with mean 0 and variance σ_ε^2. Indeed, in microeconomic panels, it is convenient (and most often relevant) to treat the cross-sectional observations as independent, identically distributed draws from the population of interest.

It is clear at this stage that, by definition of the model, the lagged endogenous variable $y_{i,t-1}$ is correlated with the random individual effects α_i. Different sets of assumptions can be made regarding the structure of the correlation of the other regressors (X, the matrix stacked version of x'_{it}) with the individual effects α_i and the error terms, ε_{it}. Sticking to the error components framework we will first assume that

---

[1] Other cases encountered in macroeconomics and finance, where the number of time periods is large, are considered in Chap. 9.

[2] Although we restrict the exposition to the case of just one lag of the dependent variable, most of the methods can be suitably adapted to the case of additional lags.

[3] Time invariant regressors do not cause any further complications per se, but they are eliminated by transformations such as the Within and first-differencing.

[4] Again, as long as the attrition is exogenous, all methods can be suitably adapted for unbalanced panels.

$$E(\alpha_i | x_{i1}, \ldots, x_{iT}) = 0, \tag{8.3}$$

and

$$E(\varepsilon_{it} | x_{i1}, \ldots, x_{iT}, y_{i,t-1}, \ldots, y_{i0}, \alpha_i) = 0,\ t = 1, \ldots, T, \tag{8.4}$$

i.e., X is strictly exogenous.

However, one can also consider that in some situations, either the former or the latter of these two assumptions, or even both, will not hold. Relaxing the former leads to the "correlated effects model" while relaxing the latter can be done by considering the regressors to be only weakly exogenous, or by assuming that they are endogenous. Both of these cases will be considered in this chapter.

A last, but important, remark has to be made at this point. Treating the number of time periods, T, as fixed has two notable implications which depart from what is generally assumed in a time-series context: (1) the stationarity assumption ($|\delta| < 1$) is not necessary; and (2) the generating process of the initial observations, y_{i0}, is important. Indeed, the correlation of the lagged endogenous variable with the disturbances of the model partly depends on this generating process. In order to highlight this dependence, let us rewrite the model, by successive backward substitution, as

$$\begin{aligned} y_{it} &= \delta^t y_{i0} + \sum_{j=0}^{t-1} \delta^j x'_{i,t-j}\beta + \frac{1-\delta^t}{1-\delta}\alpha_i + \sum_{j=0}^{t-1} \delta^j \varepsilon_{i,t-j} \\ &= \delta^t y_{i0} + \sum_{j=0}^{t-1} \delta^j x'_{i,t-j}\beta + \frac{1-\delta^t}{1-\delta}\alpha_i + \eta_{it}. \end{aligned} \tag{8.5}$$

Each observation on the endogenous variable can be written as the sum of four variables. The first one, $\delta^t y_{i0}$, depends upon the initial values; the second one on contemporaneous and past values of X; the third one, $\frac{1-\delta^t}{1-\delta}\alpha_i$, is proportional to the unobserved individual effect; and the last one can be rewritten as an autoregressive process with fixed initial values

$$\begin{aligned} \eta_{it} &= \delta \eta_{i,t-1} + \varepsilon_{it} \\ \eta_{i0} &= 0. \end{aligned} \tag{8.6}$$

The generating process of the initial observations $y_{i0}, i = 1, 2, \ldots, N$, in particular their possible correlation with the individual effects α_i, should not be ignored: it affects the asymptotic properties of many estimators. Indeed, $\mathrm{cov}(y_{i,t-1}, \alpha_i) = \delta^t \mathrm{cov}(y_{i0}, \alpha_i) + \frac{1-\delta^t}{1-\delta}\sigma_\alpha^2$, which shows that this covariance clearly depends on the assumption one makes about the initial observations. One could assume, for example, that the initial observations y_{i0} are just fixed constants or, equivalently, that they are independent of both α_i and ε_{it}. According to this rather extreme assumption, their data generating process is completely independent of that of any subsequent observation y_{it}, for $t = 1, \ldots, T$. However, since the first date of observation of the sample is often just the result of data availability, there is no real justification for such a strong assumption. Conversely, one may consider that the process under study is stationary and that all observations, whatever their date, obey the same process. If the process is purely autoregressive, this amounts to

$$y_{i0} = \frac{\alpha_i}{1-\delta} + \frac{\varepsilon_{i0}}{\sqrt{1-\delta^2}}.$$

Here one has $V(y_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}) = \sigma_\alpha^2/(1-\delta)^2 + \sigma_\varepsilon^2/(1-\delta^2)$ and $\text{cov}(y_{it}, \alpha_i|x_{i1}, x_{i2}, \ldots, x_{iT}) = \sigma_\alpha^2/(1-\delta)$, $\forall t = 0, 1, \ldots, T$. In the general case where there are exogenous regressors, one can set

$$y_{i0} = x_{i0}'\gamma^* + \frac{\alpha_i}{1-\delta} + \frac{\varepsilon_{i0}}{\sqrt{1-\delta^2}}.$$

There are many other possible assumptions that can be made here. One that implies a non-specified correlation between the initial observations and the individual effects is

$$y_{i0} = x_{i0}'\gamma^* + \phi\alpha_i + \varepsilon_{i0}.$$

Here, the conditional variance of the initial observations is $\phi^2\sigma_\alpha^2 + \sigma_\varepsilon^2$ and their covariance with the individual effects is $\phi\sigma_\alpha^2$.

Correctly specifying the generating process of the initial observations is critical when one wishes to use maximum likelihood estimation. Indeed, a misspecification in this respect would lead to inconsistent estimators in most situations (see Sect. 8.5 below as well as Nerlove (2002), or Alvarez and Arellano (2003), for a more detailed treatment of maximum likelihood estimation of dynamic models).

8.3 The Inconsistency of Traditional Estimators

For the sake of simplifying notation, let us rewrite (8.2) in more compact form, stacking over time and individuals and letting $X^* = [y_{-1}, X]$ and $\gamma = [\delta, \beta']'$

$$y = X^*\gamma + u,\ u = \alpha + \varepsilon.$$

Following Maddala (1971), we know that the usual error components model estimators belong to a general class of estimators called the "λ-class" (see Chap. 3). The λ-class estimators are computed using the OLS estimator on the transformed model

$$\left(W_N + \sqrt{\lambda}B_N\right)y = \left(W_N + \sqrt{\lambda}B_N\right)X^*\beta + \left(W_N + \sqrt{\lambda}B_N\right)u, \tag{8.7}$$

where W_N and B_N are the Within and Between matrix operators, as defined in Chap. 2 and Chap. 3.

For each $\lambda \in [0, \infty]$ one obtains estimators $\widehat{\delta}(\lambda)$ and $\widehat{\beta}(\lambda)$. This class contains all the classical estimators of the error components model such as the Within estimator (if $\lambda = 0$), the OLS estimator ($\lambda = 1$), the GLS estimator ($\lambda = \theta^2$), the Between estimator ($\lambda = \infty$), and so on. Almost all the λ-class estimators are (asymptotically in N) biased in dynamic models. To shed some light on the structure of these biases let us consider the simple AR(1) dynamic error components model ($\beta = 0$)

$$y_{it} = \delta y_{i,t-1} + \alpha_i + \varepsilon_{it}. \tag{8.8}$$

As outlined above, this equation can be written as

$$y_{it} = \delta^t y_{i0} + \frac{1-\delta^t}{1-\delta}\alpha_i + \sum_{j=0}^{t-1} \delta^j \varepsilon_{i,t-j}. \tag{8.9}$$

To facilitate the asymptotic calculations we assume that the initial observations y_{i0} are i.i.d. variables characterized by their second moments $E(y_{i0}^2)$ and by their correlation with α_i; $E(y_{i0}\alpha_i)$. Hence, the asymptotic bias of any λ-class estimator depends upon $E\left(y_{i0}^2\right)$ and $E\left(y_{i0}\alpha_i\right)$. Indeed, using the strong law of large numbers, the asymptotic limit of the above estimators, $\delta_\infty(\lambda) = \lim_{N\to\infty} \widehat{\delta}(\lambda)$, depends on the following quantities (Sevestre and Trognon, 1983, 1985)

$$\operatorname*{plim}_{N\to\infty} \frac{1}{NT} \sum\sum y_{i,t-1}(\alpha_i + \varepsilon_{it}) = \frac{1-\delta^T}{1-\delta} E(y_{i0}\alpha_i) + \frac{T-1-T\delta+\delta^T}{(1-\delta)^2}\sigma_\alpha^2$$

and

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N} \sum \bar{y}_{i,-1}(\alpha_i + \bar{\varepsilon}_i) = \frac{1-\delta^T}{1-\delta} E(y_{i0}\alpha_i) + \frac{T-1-T\delta+\delta^T}{(1-\delta)^2}(\sigma_\alpha^2 + \frac{1}{T}\sigma_\varepsilon^2).$$

The complexity of the formula does not allow us to provide the full analytical expressions of the biases for all estimators here, but Sevestre and Trognon (1983, 1985) show that $\delta_\infty(\lambda)$ is an increasing function of λ such that

$$\operatorname*{plim}_{N\to\infty} \widehat{\delta}(0) < \delta < \operatorname*{plim}_{N\to\infty} \widehat{\delta}\left(\theta^2\right) < \operatorname*{plim}_{N\to\infty} \widehat{\delta}(1) < \operatorname*{plim}_{N\to\infty} \widehat{\delta}(\infty).$$

In other words, the Within estimator of the lagged dependent variable's coefficient under-estimates its true value (Nickell, 1981), whilst the Between and OLS estimators over-estimate it. One can show that this ranking is inverted for the coefficients of the X variables as long as these variables are positively autocorrelated (and that these coefficients are positive); the opposite being true when any one of these conditions is reversed. One particular exception to this ranking is worth noting: when the initial observations are assumed to be fixed constants, or to be independent of the individual effects and the ε_{it}'s, the GLS estimator is consistent (for $N \to \infty$ but fixed T), as is the Feasible-GLS one (on the assumption that the variances σ_α^2 and σ_ε^2 have been consistently estimated). To get an idea of the biases of the OLS and Within estimators, consider, for example, a purely autoregressive model, such as (8.8), with $T = 10$, $\rho = \sigma_\alpha^2 / \left(\sigma_\alpha^2 + \sigma_\varepsilon^2\right) = 0.5$ and $\delta = 0.9$. For the Within estimator, we get

$$\operatorname*{plim}_{N\to\infty} \widehat{\delta}(0) - \delta = -0.080 \text{ if } y_{i0} = \alpha_i + \varepsilon_{i0},$$

and

$$\operatorname*{plim}_{N\to\infty}\widehat{\delta}(0)-\delta=-0.243 \text{ if } y_{i0}=\frac{\alpha_i}{1-\delta}+\frac{\varepsilon_{i0}}{\sqrt{1-\delta^2}},$$

while for the OLS estimator we have

$$\operatorname*{plim}_{N\to\infty}\widehat{\delta}(1)-\delta=0.176 \text{ and } \operatorname*{plim}_{N\to\infty}\widehat{\delta}(1)-\delta=0.095$$

for, respectively, the same two assumptions concerning y_{i0}. Clearly, the magnitude of the bias of the usual panel data estimators strongly depends on the assumptions concerning the initial observations.

However, although biased in finite T samples, the Within estimator tends to have a relatively small variance, especially when compared to consistent GMM-type estimators. In Monte Carlo studies the Within estimator appears to remain within a very tight, albeit biased, range: especially when compared to the often volatile performance of consistent estimators. Kiviet (1995) makes use of this fact combined with an approximation of the (small sample) bias of this estimator. What Kiviet (1995) suggests is to use the Within estimator, but to subtract off the approximation of its bias. In theory this yields an unbiased, or at least less biased, estimator with a very robust performance. In Monte Carlo studies, this estimator does, indeed, perform well. In practice, the drawbacks of such a procedure are twofold: the standard errors of the bias adjusted estimator are extremely complicated; and the bias adjustment term is a function of the true parameters in the model which are unknown (Kiviet (1995) suggests using an initial consistent estimator, although this conflicts with the idea of staying away from potentially volatile GMM-type estimators).

There is another simple alternative. From the above ranking of the λ-class estimators, it is clear that there exists a value $\lambda^* \in [0, \theta^2]$ such that $\lim_{N\to\infty}\widehat{\delta}(\lambda^*)=\delta$. Sevestre and Trognon (1983) have shown that

$$\lambda^*=\frac{Q(1-\rho)}{\left(\frac{1-\delta^T}{1-\delta}\frac{E(y_{i0}\alpha_i)}{\sigma^2}+Q(1-\rho+T\rho)\right)}, \tag{8.10}$$

with

$$Q=(T-1-T\delta+\delta^T)/T(1-\delta)^2,\ \rho=\sigma_\alpha^2/(\sigma_\alpha^2+\sigma_\varepsilon^2).$$

The simple method of firstly estimating λ^* by a consistent estimator $\widehat{\lambda}^*$, then δ by $\widehat{\delta}\left(\widehat{\lambda}^*\right)$ and β by $\widehat{\beta}\left(\widehat{\lambda}^*\right)$ leads to a consistent two-step estimator. Unfortunately, its finite and asymptotic distributions heavily depend on the distribution of $\widehat{\lambda}^*$: if the initial estimates of the parameters involved in λ^* are poor, the resulting $\widehat{\delta}\left(\widehat{\lambda}^*\right)$ exhibits rather unsatisfactory behavior (Sevestre and Trognon, 1990).[5]

[5] In general $\lambda^* \neq \theta^2$, which confirms the inconsistency of the GLS estimator in such a model. Nevertheless, if $E(y_{i0}\alpha_i)=0$, then $\lambda^*=\theta^2$ and this shows again that, in this particular situation, the GLS estimator is consistent.

8.4 IV and GMM Estimators

The most commonly used approach to estimating linear dynamic panel data models is undoubtedly the GMM estimator. In a general setting, recall that this method yields consistent estimators of models such as

$$y = X^*\gamma + u,$$

where X^* is used to denote that some, or all, of the regressors are correlated with the disturbances: $E(u|X^*) \neq 0$; $\text{plim}(X^{*\prime}u/NT) \neq 0$; and where these disturbances may exhibit some serial correlation or heteroskedasticity; $V(u|X^*) = \sigma^2\Omega = \Sigma$.[6] Then, assuming the existence of a set of instrumental variables, Z, satisfying in particular the absence of any (asymptotic) correlation with u and a non-null (asymptotic) correlation with the regressors, X^*, it is well known that the following instrumental variables estimator

$$\begin{aligned}\widehat{\gamma}_{IV} &= \left(X^{*\prime}Z\left(Z'Z\right)^{-1}Z'X^*\right)^{-1}X^{*\prime}Z\left(Z'Z\right)^{-1}Z'y \\ &= \left(X^{*\prime}P_ZX^*\right)^{-1}X^{*\prime}P_Zy\,,\end{aligned} \tag{8.11}$$

with $P_Z = Z\left(Z'Z\right)^{-1}Z'$, is consistent and has an asymptotic distribution

$$\sqrt{N}\left(\widehat{\gamma}-\gamma\right) \sim N\left(0, V\left(\widehat{\gamma}\right)\right),$$

where

$$\begin{aligned}V(\widehat{\gamma}) = &\left(\underset{N\to\infty}{\text{plim}}\frac{X^{*\prime}P_ZX^*}{N}\right)^{-1} \\ &\times \underset{N\to\infty}{\text{plim}}\frac{X^{*\prime}P_Z\Sigma P_ZX^*}{N} \times \left(\underset{N\to\infty}{\text{plim}}\frac{X^{*\prime}P_ZX^*}{N}\right)^{-1}.\end{aligned}$$

However, this is only efficient when the disturbances are i.i.d. (i.e., if $\Omega = I$). If they are heteroskedastic and/or serially correlated, one can improve the efficiency by resorting to the linear GMM estimator

$$\widehat{\gamma} = \left(X^{*\prime}Z\left(Z'\Sigma Z\right)^{-1}Z'X^*\right)^{-1}X^{*\prime}Z\left(Z'\Sigma Z\right)^{-1}Z'y, \tag{8.12}$$

assuming the form of Σ is known (which can happen in some cases; see Sect. 4.2.1 below). $\widehat{\gamma}$ is consistent and has an asymptotic distribution given by

$$\sqrt{N}\left(\widehat{\gamma}-\gamma\right) \sim N\left(0, V\left(\widehat{\gamma}\right)\right)$$

with

$$V(\widehat{\gamma}) = \left(\underset{N\to\infty}{\text{plim}}\frac{X^{*\prime}Z(Z'\Sigma Z)^{-1}Z'X^*}{N}\right)^{-1}.$$

[6] See Chap. 4.

If Σ is unknown, it is possible to estimate the quantity $(Z'\Sigma Z)$ required to construct the linear GMM estimator of (8.12) as

$$\widehat{Z'\Sigma Z} = \frac{1}{N}\sum_{n=1}^{N} Z_i'\widehat{u}_i\widehat{u}_i'Z_i$$

where $\widehat{u}_i$ is a vector of residuals obtained from a first step consistent estimator.

Note that all elements of X^* that are exogenous with respect to u, implicitly enter Z as their own instruments. Now, the main question is to find instruments, Z, that satisfy the above conditions given the assumptions about the disturbances including their possible correlations with the regressors. Similarly, in the discussions that follow below, any elements of X^* (or ΔX^*) that are exogenous with respect to ε and α (or $\Delta\varepsilon$)–that is the relevant error term(s) in the model–are (implicitly, or explicitly) contained as their own instruments in Z.

8.4.1 Uncorrelated Individual Effects: The Original Balestra–Nerlove Estimator and its Extensions

As outlined in the previous sections, problems arise in estimating an autoregressive (or dynamic) panel data model with unobserved effects, due to the correlation between the lagged dependent variable and these unobserved effects, α_i. In their seminal paper, Balestra and Nerlove (1966) proposed an instrumental variables approach to estimate such a model, relying on the assumption that X (or a subset of X) is independent of both α_i and ε_{it}.[7] Then these variables, and their lags, can be used as valid instruments. Following the above notation, the estimator suggested by Balestra and Nerlove (1966) is an instrumental variables estimator with the following set of instruments

$$Z_i = \begin{pmatrix} \widetilde{x}_{i0}' & \widetilde{x}_{i1}' \\ \widetilde{x}_{i1}' & \widetilde{x}_{i2}' \\ \vdots & \vdots \\ \widetilde{x}_{i,T-2}' & \widetilde{x}_{i,T-1} \\ \widetilde{x}_{i,T-1}' & \widetilde{x}_{iT}' \end{pmatrix}$$

where $\widetilde{x}_{it}$ stands for the subset of X (possibly X itself) which contains the strictly exogenous regressors and $\widetilde{x}_{i,t-1}$ their one-period lagged value.[8] In other words, in a GMM framework, the model is estimated using only the following $2 \times \widetilde{K}$ orthogonality conditions,[9] where $\widetilde{K}$ is the number of strictly exogenous regressors

[7] In fact, they assume $E(X_{i,t-\tau}\ \varepsilon_{it}) = 0,\ \tau \geq 0$.

[8] We assume that observations on y and X are available from $t = 0$ to T.

[9] In fact, one could easily go to $2 \times \widetilde{K} \times T$ instruments by considering $2 \times \widetilde{K}$ separate orthogonality conditions for each time period.

$$E\left(\widetilde{x}_{it}' u_{it}\right) = 0$$
$$\text{and } E\left(\widetilde{x}_{i,t-1}' u_{it}\right) = 0,$$

recalling that $u_{it} = \alpha_i + \varepsilon_{it}$.

One of the major drawbacks of this estimator, as of other instrumental variables estimators, is its frequent empirical imprecision, even in large samples. There are several solutions to this problem. One is to increase the number of orthogonality conditions along the lines suggested by Hausman and Taylor (1981), Amemiya and MaCurdy (1986) or Breusch, Mizon and Schmidt (1989).[10] Indeed, assuming as in Balestra and Nerlove (1966), that at least some of the regressors are both uncorrelated with the individual effects and strictly exogenous, any of the past, present or future values of these regressors (or combinations of them, such as their individual means and/or their difference from the individual means) can be used as instruments. As an example, the following set of $\widetilde{K} \times T^2$ orthogonality conditions would be valid

$$E\left(\widetilde{x}_{i\tau}' u_{it}\right) = 0,\ t = 1,\ldots,T;\ \tau = 1,\ldots,T,$$

with $\widetilde{K}$ being again, the number of strictly exogenous regressors. In the case where the regressors would still be uncorrelated with the individual effects but only weakly exogenous, then only their past values should be used as instruments.

8.4.2 Correlated Individual Effects

8.4.2.1 Model in First Differences – Instruments in Levels/First Differences

A commonly used approach to the estimation of dynamic panel data models where some (or all) elements of X are possibly correlated with the individual effects, is to write the model in first differences so that the individuals effects are, being time invariant, discarded

$$\Delta y_{it} = \delta \Delta y_{i,t-1} + \Delta x_{it}' \beta + \Delta u_{it} \tag{8.13}$$

$$= \delta \Delta y_{i,t-1} + \Delta x_{it}' \beta + \Delta \varepsilon_{it}. \tag{8.14}$$

Using previous notation one can rewrite this model in matrix form as

$$\Delta y = \Delta X^* \gamma + \Delta u$$
$$= \Delta X^* \gamma + \Delta \varepsilon$$

where the lagged endogenous variable, and possibly some other regressors, are correlated with the disturbances, both in finite samples $(E(\Delta u | \Delta X^*) \neq 0)$ and

[10] Those estimators belong to the Generalized Instrumental Variables class of estimators where the instrumental variables principle is applied to the transformed model $\Omega^{-1/2} y = \Omega^{-1/2} X\beta + \Omega^{-1/2} u$.

asymptotically $(\text{plim}\,(\Delta X^{*\prime}\Delta u/NT) \neq 0)$ and where, because of the assumed error components structure of the disturbances, $\Delta\varepsilon$ follows an $MA(1)$ process, that is[11]

$$V(\Delta u|\Delta X^*) = V(\Delta\varepsilon|\Delta X^*) = \sigma_\varepsilon^2\Omega \tag{8.15}$$

with

$$\Omega = I_N \otimes \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}.$$

Applying OLS or any other basic estimator (such as the Within or Between estimators) to this model, however, does not yield consistent estimators of the parameters, given the correlation between $y_{i,t-1}$ and $\varepsilon_{i,t-1}$. Then, one can use instrumental variables, or GMM, in this context too.

Balestra–Nerlove on First Differences

Let us first assume that the only cause of any potential endogeneity of X is through its correlation with the individual effects; that is, conditionally on the individual effects, X (or a subset, $\widetilde{X}$ thereof) is strictly exogenous. Then, these regressors or their first differences, are valid instruments for the model written in first differences. In other words, one can estimate this model using the orthogonality conditions:

$$E(\widetilde{x}_{it}\Delta u_{it}) = 0 \text{ and/or } E(\Delta\widetilde{x}_{it}\Delta u_{it}) = 0$$
$$\text{and } E(\widetilde{x}_{i,t-1}\Delta u_{it}) = 0 \text{ and/or } E(\Delta\widetilde{x}_{i,t-1}\Delta u_{it}) = 0.$$

Assuming that the variables in $\widetilde{X}$ are only weakly exogenous, only their past values up to $t-2$ at most can be used as valid instruments

$$E(\widetilde{x}_{i,t-2}\Delta u_{it}) = 0 \text{ and/or } E(\Delta\widetilde{x}_{i,t-2}\Delta u_{it}) = 0$$
$$\text{and } E(\widetilde{x}_{i,t-3}\Delta u_{it} = 0 \text{ and/or } E(\Delta\widetilde{x}_{i,t-3}\Delta u_{it}) = 0.$$

As already mentioned, the small number $(2\widetilde{K})$ of these orthogonality conditions limits the efficiency of the resulting estimators. Increasing asymptotic efficiency can be achieved by either taking into account the serial correlation of the disturbances induced by the first differencing operation and/or by using the whole sequence of $\widetilde{x}_{it}$, $t = 0,\dots,T$ as instruments, as long as they are strictly exogenous.

[11] Although more general error structure in terms of heteroskedasticity and/or serial correlation can also be allowed for (see below).

A further option was pursued by Anderson and Hsiao (1982) to get additional instruments/orthogonality conditions.

Anderson and Hsiao (1982) Estimators

As mentioned above, first differencing of the model leads to disturbances that follow an $MA(1)$ process assuming that the original ones were i.i.d. So $y_{i,t-2}$ as well as $\Delta y_{i,t-2}$ $(= y_{i,t-2} - y_{i,t-3})$, are correlated with $\Delta y_{i,t-1}$ $(= y_{i,t-2} - y_{i,t-3})$ but not with Δu_{it} $(= \Delta\varepsilon_{it} = \varepsilon_{it} - \varepsilon_{i,t-1})$, which is the disturbance of the transformed model, so they are valid instruments. In addition to $y_{i,t-2}$ and $\Delta y_{i,t-2}$ one can still use X and/or its lags, depending on whether these variables are strictly exogenous with respect to ε or not. In terms of orthogonality conditions, one can write, in the former case

$$E\left(y_{i,t-2}\Delta u_{it}\right) = 0 \text{ and/or } E\left(\Delta y_{i,t-2}\Delta u_{it}\right) = 0$$
$$\text{and } E\left(\widetilde{x}_{it}\Delta u_{it}\right) = 0 \text{ and/or } E\left(\Delta\widetilde{x}_{it}\Delta u_{it}\right) = 0$$

or, alternatively, for the latter case

$$E(y_{i,t-2}\Delta u_{it}) = 0 \text{ and/or } E(\Delta y_{i,t-2}\Delta u_{it}) = 0$$
$$\text{and } E(\widetilde{x}_{i,t-2}\Delta u_{it}) = 0 \text{ and/or } E(\Delta\widetilde{x}_{i,t-2}\Delta u_{it}) = 0.$$

The interest of these estimators is twofold. First, their implementation is very simple. Second, they do not necessarily require X to be uncorrelated with the individual effects.

However, in some instances, use of the instrument $\Delta y_{i,t-2}$ yields inefficient estimators (Arellano, 1989), suggesting that $y_{i,t-2}$ is preferable. Indeed, Anderson and Hsiao (1982) estimators in practice often appear to lead to rather erratic parameter estimates due to their low efficiency (see, for example, Sevestre, 1984 and Arellano, 1989). Several reasons may explain this. First, due to the first differencing, this method makes use of only the time variability of the observations which, at least for micro datasets, is often much less important than the between-individuals variability. Also, linked to the first differencing of the data there is the problem of low correlation between the instruments and the regressors (the so-called "weak instruments" problem). Second, due to the lags involved, these estimators effectively require the first, second and third time periods respectively to be removed, when $y_{i,t-2}$ and $\Delta y_{i,t-2}$ are used as instruments. Given the often limited time dimension of the panel, this reduces significantly the number of observations available for estimation.

Finally, as already mentioned, another reason for the often poor performance of these estimators, is the low number of instruments used. Indeed, increasing the number of instruments (in a reasonable way, see below) is known to be favourable to the behavior of the estimators. Moreover, it must also be mentioned that these estimators do not explicitly take into account the serial correlation of the disturbances which follow an $MA(1)$ process after the model has been first differenced, although it can be suitably adapted along the lines suggested above (and also below with regard to the Arellano and Bond (1991) estimator).

Arellano and Bond (1991) Estimator

Arellano and Bond (1991) proposed an estimator aimed at tackling the problems associated with the low number of instruments/orthogonality conditions and the serial correlation in the disturbances of the first differenced model.

On the first point, Arellano and Bond (1991) show that there exist many more instruments than those put forward by Anderson and Hsiao (1982). As an example, consider the case of a panel with five periods of observations, $t = 0, 1, \ldots, 4$. For each time period, the model can be written as

$$\begin{aligned} \text{for } t = 2\text{: } y_{i2} - y_{i1} &= \delta(y_{i1} - y_{i0}) + (x_{i2} - x_{i1})'\beta + \varepsilon_{i2} - \varepsilon_{i1} \\ \text{for } t = 3\text{: } y_{i3} - y_{i2} &= \delta(y_{i2} - y_{i1}) + (x_{i3} - x_{i2})'\beta + \varepsilon_{i3} - \varepsilon_{i2} \\ \text{and for } t = 4\text{: } y_{i4} - y_{i3} &= \delta(y_{i3} - y_{i2}) + (x_{i4} - x_{i3})'\beta + \varepsilon_{i4} - \varepsilon_{i3}. \end{aligned}$$

In period $t = 2$ the variable y_{i0} is a valid instrument since it is obviously correlated with $y_{i1} - y_{i0}$ but not with $\varepsilon_{i2} - \varepsilon_{i1}$ (as long as the ε_{it}'s are serially uncorrelated). Indeed, when $t = 2$, y_{i0} is nothing more than the instrument $y_{i,t-2}$ proposed by Anderson and Hsiao (1982). When $t = 3$, the instrument proposed by Anderson and Hsiao (1982) is y_{i1}. However, y_{i0} is also a valid instrument here since, given the autoregressive nature of the model, it is correlated with $y_{i2} - y_{i1}$ while, given the assumption of no serial correlation of the ε's, it is not correlated with $\varepsilon_{i3} - \varepsilon_{i2}$. This provides two instruments for estimating the model at time $t = 3$. Along the same lines, when $t = 4$, the variables y_{i0}, y_{i1} and y_{i2} are all valid instruments. The full set of instruments is given by

$$Z_i = (Z_i^{(y)}, Z_i^{(x)})$$

where

$$Z_i^{(y)} = \begin{pmatrix} y_{i0} & 0 & \ldots & \ldots & 0 & \ldots & 0 \\ 0 & y_{i0} & y_{i1} & 0 & 0 & \ldots & 0 \\ & & & \vdots & \vdots & & \\ 0 & \ldots & & & & \ldots & 0 \\ 0 & \ldots & 0 & 0 & y_{i0} & \ldots & y_{iT-2} \end{pmatrix} \tag{8.16}$$

and where the set of instruments $Z_i^{(x)}$ is defined according to the assumptions made about the exogeneity of X.[12] Although it is rather common in practice to set $Z_i^{(x)} = \Delta X$, one may augment the number of instruments based on X. For example, assuming that, conditionally on the individual effects, X is strictly exogenous (i.e., $\widetilde{X} = X$), one can use

$$Z_i^{(x)} = \begin{pmatrix} x_{i0}' & \ldots & x_{iT}' & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \ldots & 0 & x_{i0}' & \ldots & x_{iT}' & 0 & \ldots & 0 \\ 0 & \ldots & & & \ldots & 0 & \ldots & & 0 \\ 0 & \ldots & \ldots & & \ldots & 0 & x_{i0}' & \ldots & x_{iT}' \end{pmatrix} \tag{8.17}$$

[12] Along the same lines one may easily redefine the correct set of X variables that can be used as instruments depending on their correlation with α and ε.

whilst where they are only weakly exogenous, one could use

$$Z_i^{(x)} = \begin{pmatrix} x_{i0}' & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & x_{i0}' & x_{i1}' & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & & & \dots & 0 & \dots & & 0 \\ 0 & \dots & & & \dots & 0 & x_{i1}' & \dots & x_{iT-2}' \end{pmatrix}. \tag{8.18}$$

The associated orthogonality conditions can be written as

$$E(y_{i,t-\tau}\Delta\varepsilon_{it}) = 0,\ t = 2,\dots,T;\ \tau \geq 2$$

and, depending on the assumptions regarding the exogeneity of X

$$E(x_{i\tau}\Delta\varepsilon_{it}) = 0,\ t = 1,\dots,T;\ \ \tau = 1,\dots,T$$

or

$$E(x_{i,t-\tau}\Delta\varepsilon_{it}) = 0,\ t = 2,\dots,T;\ \tau \geq 2.$$

Moreover, in order to further increase efficiency, Arellano and Bond (1991) propose to account for the serial correlation of the disturbances in the first-differenced model. Indeed, one can redefine the previously defined GMM estimator, equation (8.12), except that y and X^* are replaced by their first differences and Σ by

$$\Omega = I_N \otimes \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}, \tag{8.19}$$

such that the GMM estimator writes as

$$\widehat{\gamma} = (\Delta X^{*\prime} Z \left(Z'\Omega Z\right)^{-1} Z'\Delta X^*)^{-1} \Delta X^{*\prime} Z \left(Z'\Omega Z\right)^{-1} Z'\Delta y. \tag{8.20}$$

They also suggested a variant of this estimator which is robust to heteroskedasticity. Let us denote $\widehat{\Delta u_i} = (\widehat{\Delta u_{i2}}, \widehat{\Delta u_{i3}}, \dots, \widehat{\Delta u_{iT}})'$, the vector of residuals obtained from using one of the previously presented consistent estimators–typically using (8.20) – and Ψ the unspecified variance–covariance matrix of the disturbances Δu. Then, one can consistently estimate $Z'\Psi Z$ by

$$\widehat{Z'\Psi Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i' \widehat{\Delta u_i} \widehat{\Delta u_i}' Z_i.$$

The "robust" linear GMM estimator is then

$$\begin{aligned} \widehat{\gamma}_r = &\left(\Delta X^{*\prime} Z (\widehat{Z'\Psi Z})^{-1} Z'\Delta X^*\right)^{-1} \\ &\times \Delta X^{*\prime} Z (\widehat{Z'\Psi Z})^{-1} Z'\Delta y. \end{aligned} \tag{8.21}$$

There is evidence that the standard errors of the robust two-step GMM variant of this estimator are unreliable unless N is very large. Windmeijer (2005) however, suggests a method for correcting this small-sample bias (see Sect. 8.6).

Unfortunately, as already mentioned regarding the Anderson and Hsiao (1982) estimators, those suggested by Arellano and Bond (1991) also often suffer from a lack of precision due to the first differencing of the model and the lack of correlation of the instruments with the regressors. This may sometimes result in rather erratic parameter estimates. Some alternatives do exist though, to circumvent this problem.

8.4.2.2 Model in Levels – Instruments in First Differences

As already mentioned in Chap. 4, there are several ways to tackle the problem of unobserved effects being correlated with explanatory variables. As sketched out above, the most common one, the Within transformation, does not work in the context of dynamic models, nor does OLS on a first differenced model. Indeed, first differencing creates $MA(1)$ disturbances for the transformed model, thus inducing a correlation with the lagged dependent variable.

However, assuming that the covariance between the regressors X and the individual effects is constant over time, the first differences of those regressors are uncorrelated with the unobserved effects. This makes them valid instruments as long as they are also exogenous with respect to ε_{it}. Indeed, assuming that

$$\operatorname{cov}(x_{it}, \alpha_i) = \Gamma_i \,, \quad \forall t$$

and that, conditionally on the unobserved effects, X is strictly exogenous, one can make use of the following instruments

$$Z_i^{(\Delta x)} = \begin{pmatrix} \Delta x_{i1}' & \dots & \Delta x_{iT}' & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & \Delta x_{i1}' & \dots & \Delta x_{iT}' & 0 & \dots & 0 \\ 0 & \dots & & & \dots & 0 & \dots & & 0 \\ 0 & \dots & & & \dots & 0 & \Delta x_{i1}' & \dots & \Delta x_{iT}' \end{pmatrix} \tag{8.22}$$

while, if they are only weakly exogenous, we can use

$$Z_i^{(\Delta x)} = \begin{pmatrix} \Delta x_{i2}' & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & \Delta x_{i2}' & \Delta x_{i3}' & 0 & \vdots & & 0 \\ \vdots & & & \ddots & \ddots & \vdots & \\ 0 & \dots & 0 & 0 & \Delta x_{i2}' & \cdots & \Delta x_{i,T-1}' \end{pmatrix}.$$

This corresponds to the following orthogonality conditions

$$E(\Delta x_{i\tau}' u_{it}) = 0, \ \tau = 1, \dots, T; \ \ t = 1, \dots, T$$

if the regressors are strictly exogenous, and

$$E(\Delta x'_{i,t-j} u_{it}) = 0,\ j = 1, \ldots,\ t-1,\ t = 2, \ldots, T$$

if those regressors are only weakly exogenous.

Moreover, under the assumption that the covariance between the lagged endogenous variable and the individual effects is also constant over t (which amounts to assuming stationarity of y_{it}), lagged values of the first difference of that lagged endogenous variable are also valid instruments as they are in this case uncorrelated with the disturbances.[13]

Indeed, as an example, let us consider again the simple case of a panel with five periods of observations ($t = 0, 1, \ldots, 4$). The model can be written as

$$\begin{aligned} \text{for } t = 2 : y_{i2} &= \delta y_{i1} + x'_{i2}\beta + \alpha_i + \varepsilon_{i2} \\ \text{for } t = 3 : y_{i3} &= \delta y_{i2} + x'_{i3}\beta + \alpha_i + \varepsilon_{i3} \\ \text{and for } t = 4 : y_{i4} &= \delta y_{i3} + x'_{i4}\beta + \alpha_i + \varepsilon_{i4}. \end{aligned}$$

In period $t = 2$ the variable $y_{i1} - y_{i0}$ is a valid instrument since it is obviously correlated with y_{i1} but not with ε_{i2} as long as the ε_{it}'s are serially uncorrelated and that $\text{cov}(y_{i1} - y_{i0}, \alpha_i) = 0$ as long as the stationarity assumption is fulfilled since in that case, $\text{cov}(y_{i1}, \alpha_i) = \text{cov}(y_{i0}, \alpha_i) = \sigma_\alpha^2/(1-\delta)$. Along the same lines, when $t = 3$, $y_{i2} - y_{i1}$ is a valid instrument while $y_{i1} - y_{i0}$ also remains valid. This provides two instruments for estimating the model at time $t = 3$.With $t = 4$ the variables $y_{i1} - y_{i0}, y_{i2} - y_{i1}$ and $y_{i3} - y_{i2}$ are now all valid instruments, such that one could also consider the following set of instruments

$$Z_i^y = \begin{pmatrix} \Delta y_{i1} & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & \Delta y_{i1} & \Delta y_{i2} & 0 & \vdots & & 0 \\ \vdots & & & \ddots & \ddots & \vdots & \\ 0 & \ldots & 0 & 0 & \Delta y_{i1} & \cdots & \Delta y_{i,T-1} \end{pmatrix}.$$

That is

$$E(\Delta y_{i,t-1} u_{it}) = 0,\ t = 2, \ldots, T.$$

8.4.2.3 Other Transformations for Models with Correlated Individual Effects

Arellano and Bover Estimator

A drawback associated with first differencing of the model is the induced serial correlation in the disturbances of the transformed model. This is why Arellano and Bover (1995) suggested a transformation of the model that discards the individual effects without inducing any serial correlation in the remaining disturbances. This transformation, called *forward orthogonal deviation*, is defined by the $(T-1) \times T$ matrix

[13] This argument closely follows that of Blundell and Bond (1998) discussed below.

$$H^* = \text{diag}[(T-1)/T,\ (T-2)/(T-1),\ (T-3)/(T-2), \ldots, 1/2] \times H^+$$

with

$$H^+ = \begin{pmatrix} 1 & 1/(1-T) & 1/(1-T) & \cdots & 1/(1-T) & 1/(1-T) & 1/(1-T) \\ 0 & 1 & 1/(2-T) & \cdots & 1/(2-T) & 1/(2-T) & 1/(2-T) \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1/2 & -1/2 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix}.$$

Applying this transformation to the vector u_i of the disturbances ($u_{it} = \alpha_i + \varepsilon_{it}$) for the ith individual will result in a transformed vector u_i^* of size $(T-1)$ defined as

$$u_i^* = H^* u_i$$

with typical elements

$$\begin{aligned} u_{it}^* &= d_t \times (u_{it} - \frac{1}{T-t}(u_{i,t+1} + u_{i,t+2} + \ldots + u_{iT})) \\ &= d_t \times (\varepsilon_{it} - \frac{1}{T-t}(\varepsilon_{i,t+1} + \varepsilon_{i,t+2} + \ldots + \varepsilon_{iT})) \end{aligned}$$

where $d_t = (T-t)/(T-t+1)$.

The interesting result here is that we have $V(u_i^*) = \sigma^2 I_{T-1}$, i.e., the transformed disturbance vector does not include the individual effects anymore and is, moreover, serially uncorrelated. So, instrumental variable estimators of the H^* transformed model will not be subject to any loss of efficiency because of ignored serial correlation. In particular, any set of predetermined variables (be they lagged values of y or X) will constitute valid instruments.

Explicitly, the estimator suggested by Arellano and Bover (1995) first involves transforming the system of T equations using the nonsingular transformation given by H_i

$$H_i = \begin{bmatrix} H^* \\ e_T/T \end{bmatrix} \tag{8.23}$$

where e_T is a (column) vector of ones of size T. H^* can be defined as above, or indeed by any $(T-1) \times T$ matrix of rank $(T-1)$ such that $H^* e_T = 0 : H^*$ for example, could also be the first $(T-1)$ rows of the Within group operator or the first difference operator, although interestingly, the estimator is invariant to the specific choice of H^*. As the first $(T-1)$ transformed errors

$$u_i^+ = H_i u_i = \begin{bmatrix} H^* u_i \\ \overline{u}_i \end{bmatrix}, \tag{8.24}$$

are free of α_i, all exogenous variables are valid instruments for these first equations. Moreover, assuming serial independence of the disturbance terms ε_{it}, along the lines of the Arellano and Bond (1991) estimator, the series $(y_{i0}, y_{i1}, \ldots, y_{i,t-1})$ is also a

valid instrument. This assumption however, requires more structure for H^*, which now additionally has to be upper triangular (Arellano and Bover, 1995). This defines the matrix of valid instruments to be

$$Z_i^{ABov} = \begin{bmatrix} (X_i', y_{i0}) & & & 0 \\ & (X_i', y_{i0}, y_{i1}) & & \\ & & \ddots & \\ & & & (X_i', y_{i0}, \ldots, y_{i,T-1}) \\ 0 & & & X_i' \end{bmatrix}, \tag{8.25}$$

where $X_i' = (x_{i1}', \ldots, x_{iT}')$ as long as X is strictly exogenous. Stacking the Z_i's and letting $H = I_N \otimes H_i$, the estimator is obtained by estimating the transformed model

$$Z'Hy = Z'H\widetilde{X}\gamma + Z'Hu \tag{8.26}$$

by GLS using $V(Z'Hu) = Z'H\Sigma H'Z$.

As with previous estimators, the covariance of the transformed system $\Sigma_H = H\Sigma H'$ must be estimated from residuals obtained from a preliminary consistent estimator of $H_i u_i = \widehat{u}_i^+$. One option is Arellano and Bover (1995)

$$\widehat{\Sigma}_H = \frac{1}{N}\sum_i \widehat{u}_i^+ \widehat{u}_i^{+\prime}, \tag{8.27}$$

which is an unrestricted estimator of Σ_H. The restricted estimator under the usual assumptions of the error components model is

$$\widehat{\Sigma}_H = H\widehat{\Sigma}H' \tag{8.28}$$

where $\widehat{\Sigma} = I_N \otimes \left(\widehat{\sigma}_\alpha^2 J_T + \widehat{\sigma}_\varepsilon^2 I_T\right)$ and $\widehat{\sigma}_\alpha^2$ and $\widehat{\sigma}_\varepsilon^2$ are consistent estimators of σ_α^2 and σ_ε^2.

8.4.2.4 Wansbeek–Bekker Estimator

The approach suggested by Wansbeek and Bekker (1996) extends those of Anderson and Hsiao (1982), Arellano (1989) and Arellano and Bond (1991), for example, such that now both lags and leads (and linear combinations) of the dependent variable are included in the instrument set. By defining the variable y from period $t = 1$ to $t = T$, the estimator considers linear functions of y_+ as instruments, where y_+ is the stacked vector of observations defined from $t = 0$ to $t = T$ for each individual. The linear functions are defined by a $((T+1) \times T)$ matrix A_i, which yields $A_i' y_+$ as the full instrument set. Important restrictions are imposed on A_i such that

$$A_i e_T = 0 \text{ and } E\left(y_+' A_i \varepsilon\right) = \text{trace} A_i E\left(\varepsilon_i y_+'\right) = 0 \tag{8.29}$$

which respectively ensure elimination of the individual effects and consistency of the estimator (ε_i is the $T \times 1$ vector of disturbances ε_{it} for individual i).

These conditions for A_i define its structure such that its rows must sum to zero, as must each of its lowest T quasi-diagonal elements (in particular, the lower left element is zero). A_i is unspecified apart from the restrictions of (8.29); the estimator's variance is a function of A_i. The optimal value of A_i can be found by constrained optimisation such that the estimator's (scalar) variance is minimised whilst ensuring that the appropriate restrictions hold (Wansbeek and Bekker, 1996). However, this is done under the simplifying assumption that $\beta = 0$ although the estimator can be "operationalised" to allow for exogenous variables as well (Harris and Mátyás, 2000). In this instance, the full operational instrument set is defined as

$$Z^{WB} = \left(A'y_+, X\right). \tag{8.30}$$

With known A_i and hence known Z, the (operational) estimator is a straightforward application of the linear GMM estimator, using $\sigma_u^2 \left(Z'Z\right)$ as an approximation to the variance of $Z'u$. The asymptotic variance is given by

$$\sigma_u^2 \left(\underset{N\to\infty}{\text{plim}} \frac{1}{N} \left(X^{*\prime} P_Z X^*\right)^{-1} \right), \tag{8.31}$$

where $P_Z = Z\left(Z'Z\right)^{-1} Z'$, which, from (8.30) is a function of A_i. The optimal choice of A_i is that which minimises (8.31), such that A_i conforms with its appropriate restrictions. If one is only interested in the variance of the parameter vector (and not covariances of particular elements of it), the optimal estimator can be obtained by constrained optimisation, where A_i is that which minimises the trace of (8.31), treating σ_u^2 as a constant, subject to the restrictions given by (8.29).

The list of valid instruments can also be expanded to include not only $A'y_+$, but also $A'X_+$ for example, such that

$$Z^{WB^+} = \left(A'y_+, A'X_+, X\right). \tag{8.32}$$

Again, the instrument set for these estimators can suitably be adopted for different assumptions regarding the exogeneity of the elements of X. Despite some evidence on the good performance of these estimators (Wansbeek and Bekker, 1996; Harris and Mátyás, 2000, 2004), they are not yet commonly used in practice.

8.4.2.5 Combining Levels and Differences: The Blundell and Bond System Estimator

Over the last decade the practice of estimating dynamic panel data models has mainly consisted of using the GMM estimator suggested by Arellano and Bond (1991) i.e., first differencing the model and using lagged levels of the endogenous explanatory variable as instruments. From an empirical point of view though, the resulting parameter estimates were often unsatisfactory: imprecise and not very robust (for example, to a slight change in the instrument set). Blundell and Bond (1998) have shown that one likely explanation for this is the lack of correlation between the instruments (considered to be lagged values of the endogenous variable (e.g., $y_{i,t-2}$

and so on), and the regressors in the model once it has been first differenced. This is the now well-known *weak instruments* problem.

Assume, for the sake of simplicity, that the sample consists of three periods $(t = 0, 1, 2)$, and that the model is a purely autoregressive one. The only valid instrument is $y_{i,t-2}$ and the appropriate orthogonality condition is

$$E\left[y_{i0}\left(\Delta y_{i2} - \delta \Delta y_{i1}\right)\right] = 0.$$

The question arises of the correlation between the regressor Δy_{i1} and the instrument y_{i0}. Given the specification of the model, one can write

$$y_{i1} = \delta y_{i0} + \alpha_i + \varepsilon_{i1}$$

and

$$\begin{aligned} \Delta y_{i1} &= (\delta - 1) y_{i0} + \alpha_i + \varepsilon_{i1} \\ &= \pi y_{i0} + \alpha_i + \varepsilon_{i1}. \end{aligned}$$

Blundell and Bond (1998) show that

$$\operatorname*{plim}_{N\to\infty} \widehat{\pi} = (\delta - 1)\, \frac{q}{(\sigma_\alpha^2/\sigma_\varepsilon^2) + q} \text{ with } q = \frac{(1-\delta)^2}{1-\delta^2}.$$

In other words, when $\delta \to 1$, or when the ratio $\sigma_\alpha^2/\sigma_\varepsilon^2 \to \infty$, the correlation between the instrument and the regressor tends to 0, which induces the erratic behavior of the estimator. Blundell and Bond (1998) therefore suggest adding supplementary orthogonality conditions, based on the additional assumption of "quasi-stationarity" of y_{it}. This assumption amounts to considering that the initial observations y_{i0} are generated according to

$$y_{i0} = x_{i0}'\varphi + \frac{\alpha_i}{1-\delta} + \varepsilon_{i0}.$$

Indeed, under this assumption, one can write

$$E\left(\Delta y_{i1}\, \varepsilon_{i2}\right) = 0$$

and, with a larger number of periods,

$$E\left(\Delta y_{i,t-1}\, \varepsilon_{it}\right) = 0,\ t = 2, 3, \ldots, T.$$

One can then instrument the model in levels by lagged first differences of y_{it}, following the suggestion made by Arellano and Bover (1995) for models with correlated individual effects. Blundell and Bond (1998) then suggest a "system GMM" estimator that consists of stacking the Arellano and Bond (1991) orthogonality conditions with the above ones. In other words, one stacks the model "in levels" and that in "first differences"

$$\begin{pmatrix} \Delta y \\ y \end{pmatrix} = \delta \begin{pmatrix} \Delta y_{-1} \\ y_{-1} \end{pmatrix} + \begin{pmatrix} \Delta X \\ X \end{pmatrix} \beta + \begin{pmatrix} \Delta \varepsilon \\ \alpha + \varepsilon \end{pmatrix},$$

and estimate this system using GMM with the following set of instruments $Z = Z^{\Delta}$ and Z^{l}, where the former correspond to the instruments for the model in first differences, and the latter to those associated with the model in levels, such that

$$Z_i^l = \begin{pmatrix} \Delta y_{i2} & 0 & \dots & 0 \\ 0 & \Delta y_{i3} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \Delta y_{i,T-1} \end{pmatrix}$$

and in full

$$Z_i = \begin{pmatrix} Z_i^{\Delta} & 0 & 0 & \dots & 0 \\ 0 & \Delta y_{i2} & 0 & \dots & \dots \\ \vdots & 0 & \Delta y_{i3} & 0 & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \dots & 0 & \Delta y_{i,T-1} \end{pmatrix}$$

Blundell and Bond (1998) provide some Monte-Carlo simulation results showing that this procedure leads to a more efficient and robust estimator. Of course, one could further consider extensions to such an approach: any of the valid instruments previously described for the model in first differences are valid candidates for Z^{Δ} (although Blundell and Bond (1998) suggest using those given in Arellano and Bond (1991)); and, depending upon the assumptions one is prepared to make regarding the exogeneity of the X variables, further moment conditions based on these, as suggested above, could also be used to further augment the instrument set.

8.4.2.6 Using Further Orthogonality Conditions: Ahn and Schmidt Estimator

It is generally considered that increasing the number of orthogonality conditions is a way to improve the asymptotic efficiency of an estimator; an argument used by Arellano and Bond (1991), for example. Along the same lines, Ahn and Schmidt (1995, 1997, 1999) have shown that one can deduce supplementary orthogonality conditions from some of the assumptions that define the dynamic error components model.

Indeed, the assumption of absence of serial correlation of the ε_{it} disturbances implies the following orthogonality conditions

$$E\left(\varepsilon_{iT}\Delta\varepsilon_{it}\right) = 0,\ t = 2,\dots,T-1$$

that is

$$E\left[\left(y_{iT} - \delta y_{i,T-1} - x_{iT}'\beta\right)\left(\Delta y_{it} - \delta\Delta y_{i,t-1} - \Delta x_{it}'\beta\right)\right] = 0,\ t = 2,\dots,T-1.$$

Moreover, the homoskedasticity of ε_{it} allows one to write

$$E\left(y_{i,t-2}\Delta\varepsilon_{i,t-1} - y_{i,t-1}\Delta\varepsilon_{it}\right) = 0,\ t = 4,\ldots,T$$

that is

$$E[y_{i,t-2}(\Delta y_{i,t-1} - \delta\Delta y_{i,t-2} - \Delta X'_{i,t-1}\beta) - y_{i,t-1}(\Delta y_{it} - \delta\Delta y_{i,t-1} - \Delta X'_{it}\beta)] = 0,$$
$$t = 3,\ldots,T.$$

Estimation can now be undertaken using non-linear GMM techniques. However, the asymptotic efficiency gain may not prove to be very important, and moreover, must be weighed against the increased complexity of the estimation procedure.

8.4.3 Some Monte Carlo Evidence

An extensive amount of work has been undertaken analysing the small sample performance of the GMM-type estimators (see, for example, Arellano and Bond, 1991, Arellano and Bover, 1995, Kiviet, 1995, Judson and Owen, 1999, Harris and Mátyás, 2004). With such a vast array of experimental evidence, results, of course, vary quite dramatically depending on sample sizes (in N, T and the number of Monte Carlo replications), parameter settings and the assumed data generating process (for example, whether the model is strictly a simple autoregressive one, or not).

Notwithstanding these comments, in the majority of the studies simple IV estimators that have the same number of instruments as endogenous variables, fare poorly due to their lack of finite moments (Kinal, 1980). Typically, Arellano's (1989) estimator tends to have much better performance than that of Anderson and Hsiao (1982), as do the Arellano and Bond (1991) one(s), although often their relative performance is quite close. A clear message is that in small T panels, GMM-estimators using a large number of moment conditions (for example, the Arellano and Bond (1991), Arellano and Bover (1995) and non-linear GMM estimators) can all suffer, to a significant extent, from the resulting small-sample bias. Therefore, if practitioners are using such estimators they should attempt to limit the number of conditions so used. An estimator which appears to have an extremely robust performance is the Balestra and Nerlove (1966) estimator, which is, moreover, very simple to construct. Not only does this estimator perform well across numerous settings (parameters, sample sizes) but has also been shown, in a Monte Carlo setting, to perform very well when the true data generating process has been seriously misspecified (Harris and Mátyás, 1996). Also evident from the numerous Monte Carlo results is that as $\delta \to 1$, simple OLS (or FGLS) is the preferred option.

8.5 The Maximum Likelihood Estimator[14]

An alternative to GMM estimation is to use maximum likelihood techniques, assuming that both the individual effects α_i and the idiosyncratic disturbances ε_{it} are normally distributed. A simple way to estimate the model might be to consider the likelihood of the observed sample for $t = 1,\ldots,T$, conditional on the initial observations y_{i0}

$$L_{NT}(\delta,\beta,\sigma_\alpha^2,\sigma_\varepsilon^2|y_0,X) = -\frac{NT}{2}\ln 2\pi - \frac{N}{2}\ln\det(\Omega) - \frac{1}{2}\sum_i u_i'\Omega^{-1}u_i$$

with

$$u_i' = (y_{i1} - \delta y_{i0} - x_{i1}'\beta,\ldots,y_{iT} - \delta y_{i,T-1} - x_{iT}'\beta)$$
$$\Sigma = \sigma_\varepsilon^2[W_N + (1/\theta)B_N].$$

This was first explored in the seminal paper by Balestra and Nerlove (1966), with a conditional likelihood function in which the initial observations were assumed to be fixed; the main advantage of this assumption being that this leads to GLS-like estimators, which are quite easy to compute. However, such ML estimators are, for a wide range of combinations of the parameters, equal to the OLS estimator and hence they are not consistent (Trognon, 1978).

This important drawback does not occur when the likelihood function takes into account the density function of the first observations, i.e., when the likelihood function is unconditional (Bhargava and Sargan, 1983). To illustrate the unconditional ML estimator, let us consider the following model

$$y_{it} = \delta y_{i,t-1} + x_{i1}^{*\prime}\beta + z_i^*\gamma^* + \alpha_i + \varepsilon_{it} \tag{8.33}$$

where, along the lines suggested by Bhargava and Sargan (1983), Chamberlain (1982) or Blundell and Smith (1991), the initial values are assumed to be defined by

$$y_{i0} = \varphi z_i + u_{i0}\,. \tag{8.34}$$

It is convenient to decompose the specific effect α_i as in the following regression

$$\alpha_i = \psi u_{i0} + \zeta_i\,,$$

where ζ_i is independent of u_{i0}. In this model, $(u_{i0},\zeta_i,\varepsilon_{i1},\ldots,\varepsilon_{iT})$ are i.i.d. $N(0,$ $\mathrm{diag}(\sigma_u^2,\sigma_\zeta^2,\sigma_\varepsilon^2 e_T'))$ and the log-likelihood function is

[14] For a detailed exposition of the application of maximum likelihood techniques to dynamic error components models, see Nerlove (2002).

$$L_{NT}(\delta,\beta,\gamma,\psi,\varphi,\sigma_u^2,\sigma_\zeta^2,\sigma_\varepsilon^2) = -\frac{N(T+1)}{2}\ln\ 2\pi - \frac{N}{2}\ln\ \sigma_u^2 - \frac{1}{2\sigma_u^2}\sum_i u_{i0}^2 - \frac{N}{2}\ln\det(\Omega) - \frac{1}{2}\sum_i u_i^{*\prime}\Omega^{-1}u_i^*$$

with

$$u_i^{*\prime} = (y_{i1} - \delta y_{i0} - \beta x_{i1} - \gamma z_i - \varphi u_{i0}, \ldots, y_{iT} - \delta y_{i,T-1} - \beta x_{iT} - \gamma z_i - \varphi u_{i0})$$
$$u_{i0} = y_{i0} - \varphi z_i.$$

The first order conditions are then as follows

$$a.\quad \frac{\partial L}{\partial \delta} = \frac{1}{\sigma_\varepsilon^2}\sum_i y_{i,-1}W_N u_i^* + \frac{1}{(\sigma_\varepsilon^2 + T\sigma_\zeta^2)}\sum_i y_{i,-1}'B_N u_i^* = 0 \tag{8.35}$$

$$b.\quad \frac{\partial L}{\partial \beta} = \frac{1}{\sigma_\varepsilon^2}\sum_i x_i'W_N u_i^* + \frac{1}{(\sigma_\varepsilon^2 + T\sigma_\zeta^2)}\sum_i x_i'B_N u_i^* = 0$$

$$c.\quad \frac{\partial L}{\partial \gamma} = \frac{1}{(\sigma_\varepsilon^2 + T\sigma_\zeta^2)}\sum_i z_i e_T' u_i^* = 0$$

$$d.\quad \frac{\partial L}{\partial \varphi} = \frac{1}{(\sigma_\varepsilon^2 + T\sigma_\zeta^2)}\sum_i u_{i0} e_T'\ u_i^* = 0$$

$$e.\quad \frac{\partial L}{\partial \psi} = \frac{1}{(\sigma_\varepsilon^2 + T\sigma_\zeta^2)}\sum_i \varphi z_i e_T'\ u_i^* + \frac{1}{\sigma_u^2}\sum_i z_i\ u_{i0} = 0$$

$$f.\quad \frac{\partial L}{\partial \sigma_\varepsilon^2} = -\frac{N(T-1)}{2\sigma_\varepsilon^2} + \frac{1}{2\sigma_\varepsilon^4}\sum_i u_i^{*\prime}W_N u_i^* = 0$$

$$g.\quad \frac{\partial L}{\partial (\sigma_\varepsilon^2 + T\sigma_\zeta^2)} = -\frac{N}{(\sigma_\varepsilon^2 + T\sigma_\zeta^2)} + \frac{1}{(\sigma_\varepsilon^2 + T\sigma_\zeta^2)^2}\sum_i u_i^{*\prime}B_N\ u_i^* = 0$$

$$h.\quad \frac{\partial L}{\partial \sigma_u^2} = -\frac{N}{2\sigma_u^2} + \frac{1}{2\sigma_4^u}\sum_i u_{i0}^2 = 0\,.$$

This set of equations implies that the ML estimators of φ and σ_u^2 are OLS estimators on equation (8.34). If $\widehat{u}_{i0}$ is the residual of this equation, then the other ML estimators are the solutions of (8.35), where u_{i0} has been replaced by $\widehat{u}_{i0}$ in u_i^*. In other words, the ML estimators of the model given by (8.33) and (8.34), are the combination of OLS estimation of (8.34) followed by ML estimation of (8.33), where the unobservable u_{i0} is replaced by its estimator $\widehat{u}_{i0}$.

In the case where the initial observations are defined as

$$y_{i0} = \varphi z_i + \kappa x_{i0} + u_{i0} \tag{8.36}$$

but the model for other periods as given in (8.33) is unchanged, the variable x_{i0} does not enter the autoregressive equation and the previous simplification disappears. Nevertheless here, Sevestre and Trognon (1990) suggest an estimation technique

equivalent to ML estimation based on the following extended autoregressive model

$$y_{it} = \delta y_{i,t-1} + x_{it}'\beta + \gamma z_i + \rho x_{i0} + \psi u_{i0} + \zeta_i + \varepsilon_{it}.$$

The ML estimator of δ, β, γ, ψ, σ_ζ^2, σ_u^2, σ_ε^2 on this auxiliary model are asymptotically equivalent to the ML estimator of the true model. If $\widehat{\kappa}$ is the OLS estimator of κ on (8.36), a more efficient estimator is given by $\widehat{\kappa}^* = \widehat{\kappa} + \widehat{\rho}^*/\widehat{\psi}^*$, which is asymptotically efficient if $\widehat{\rho}^*$ and $\widehat{\psi}^*$ are the ML estimators of ρ and ψ in the auxiliary model (see Sevestre and Trognon, 1990 and Blundell and Smith, 1991).

It appears that the unconditional ML estimator can be worked out in a very simple way. Obviously, in the case when the disturbances are normal, such an ML procedure yields the asymptotically most efficient estimator, but this good behavior is also apparent in small samples too as evidenced by simulation studies (Sevestre and Trognon, 1990).

8.6 Testing in Dynamic Models

8.6.1 Testing the Validity of Instruments

Many potential candidates have been presented so far that consistently estimate a linear dynamic panel data model. Of course, an obvious question is *which one to use?,* as in practice different choices can lead to vastly different parameter estimates and inference (Lee et al., 1998). One can base judgment on the available Monte Carlo evidence (see Sect. 4.3), but this is typically inconclusive. The concept of a "good" instrument is twofold: exogeneity and relevance (one might also be concerned with efficiency in practice, with computational complexity).

The assumption about the absence of any (asymptotic) correlation between the instrumental variables and the disturbances is commonly tested using the Sargan (Hansen) test (Sargan, 1958 and Hansen, 1982). This is dealt with in Chap. 4 and as such, we do not dwell on this issue here. An issue to bear in mind here is that for this (these) test(s) to be valid, any heteroskedasticity and/or serial correlation present, must have been taken into account in estimation.

The other aspect of the choice of "good" instruments lies in the strength of the correlation between the endogenous regressors and the instruments. Indeed, the concept of relevance and of weak instruments in the dynamic linear panel setting was considered by Blundell and Bond, (1998), who show that a small correlation results in erratic parameter estimates. Moreover, there is a significant amount of recent literature highlighting the deleterious effects of weak instruments in general (see, for example Wang and Zivot, 1998, Woglom, 2001 and Hahn and Hausman, 2002).

Almost inevitably, the discussion of instrument relevance has traditionally focussed on the correlation between the endogenous regressor(s) and the instrument(s). This is the approach followed in Bound et al. (1995) and Shea (1997). Thus

one could potentially choose across estimators, based on R^2 and partial R^2 statistics. Such an approach is taken further by Poskitt and Skeels (2004) who, instead, develop a procedure based on a measure of lack of correlation. This approach appears preferable, as it is possible to develop the sampling distribution of the proffered measure, such that standard inferential procedures can be utilised. Moreover, the Poskitt and Skeels (2004) test statistic is readily computed using a set of auxiliary regressions and can be extended to the case where there are multiple endogenous regressors. This procedure, which has yet to be explicitly extended for use within a panel data setting, could thus be used to statistically test for the presence of weak instruments and, in conjunction with the Sargan test, be used to determine the most appropriate IV estimator. Another way to determine the relevance of the instruments is to compute canonical correlations (Mairesse et al., 1999).

8.6.2 Testing for Unobserved Effects

As shown in previous chapters, there are many ways to test for the presence of unobserved effects. In the case of the dynamic panel model, it is possible to use Hausman's (1978) test statistic. If the regressors, except for the lagged endogenous variable, are strictly exogenous and the ε_{it} disturbances are homoskedastic and serially uncorrelated, then in the absence of any individual effects, i.e., under H_0, the OLS estimator is consistent and asymptotically efficient. However, as shown, it is inconsistent when such effects exist. On the other hand, we have presented numerous estimators which are consistent whether there are individual effects or not (denote these generically $\widehat{\gamma}^*$). Thus in order to test for the presence of individual effects, one may compute the following statistic[15]

$$Q_H = (\widehat{\gamma}^* - \widehat{\gamma}_{OLS})' \left[\widehat{V}(\widehat{\gamma}^*) - \widehat{V}(\widehat{\gamma}_{\mathrm{OLS}})\right]^{-1} (\widehat{\gamma}^* - \widehat{\gamma}_{\mathrm{OLS}})$$

Asymptotically under the null hypothesis, $Q_H \sim \chi^2_{\dim(\gamma)}$ and the test rejects for large values: OLS is inconsistent and one must include unobserved effects in the model and apply consistent techniques. On the other hand, under H_0, OLS can be safely used.

However, it is rather unlikely in practice that all of the assumptions ensuring the optimality of OLS in the absence of individual effects, are satisfied. When the disturbances are heteroskedastic, one would use (Feasible) GLS instead of OLS as a basis of comparison with GMM estimates. If the ε_{it} disturbances are serially correlated, thus inducing a correlation with the lagged endogenous variable, one should compare two different GMM estimators. One estimator would be consistent only in the absence of individual effects (e.g., in the case where the ε_{it} generating process is an $MA(q)$, one might think of using lags of order $q+1$ or more of the endoge-

[15] In cases where the difference in variances appears to be non-invertible, one may resort to the generalised inverse.

nous variables as instruments; indeed, this is consistent under H_0 but not under the alternative). The other estimator might make use of strictly and doubly exogenous variables as instruments (i.e., variables that are exogenous both with respect to the individual effects and the ε_{it} disturbances).

8.6.3 Testing for the Absence of Serial Correlation in ε

The presence of serial correlation in ε_{it} will typically invalidate the use of lagged values (and first differences of such) of the endogenous variable as instruments. So in these circumstances, it is crucial to test for such serial correlation. Probably the most widely used test here is that proposed by Arellano and Bond (1991), which is based upon the model estimated in first differences. Let $\widehat{\Delta\varepsilon}$ be the vector of residuals from the model in first differences, $\widehat{\Delta\varepsilon}_{-2}$ its second lag value, and $\widehat{\Delta\varepsilon}^{*}$ the reduction of the vector $\widehat{\Delta\varepsilon}$ allowing computation of the product $\widehat{\Delta\varepsilon}'_{-2}\widehat{\Delta\varepsilon}^{*}$.[16] The test statistic is

$$m_2 = \frac{\widehat{\Delta\varepsilon}'_{-2}\widehat{\Delta\varepsilon}^{*}}{\widehat{\xi}^{1/2}}$$

where

$$\begin{aligned}\widehat{\xi}^{1/2} = &\sum_i \widehat{\Delta\varepsilon}'_{i,-2}\widehat{\Delta\varepsilon}^{*}_i\widehat{\Delta\varepsilon}^{*\prime}_i\widehat{\Delta\varepsilon}_{i,-2} \\ &- 2\widehat{\Delta\varepsilon}'_{-2}X^{*}_{*}\left[X^{*\prime}Z'\left(\widehat{\sigma^2 Z'\Omega Z}\right)^{-1}Z'X^{*}\right]^{-1} \\ &\times\ X'Z\left(\widehat{\sigma^2 Z'\Omega Z}\right)^{-1}\sum_i Z_i'\widehat{\Delta\varepsilon}_i\widehat{\Delta\varepsilon}^{*\prime}\widehat{\Delta\varepsilon}_{i,-2} \\ &+ \widehat{\Delta\varepsilon}'_{-2}X^{*}_{*}V(\widehat{\gamma})X^{*\prime}_{*}\widehat{\Delta\varepsilon}_{-2}\,,\end{aligned}$$

with X^{*}_{*} similarly defined as the reduction of the regressors matrix $X^{*} = [y_{-1}, X]$ and where, as before, $\widehat{\gamma} = [\widehat{\delta}, \widehat{\beta}\,']'$.

This test provides a measure of the importance of serial correlation of order 2 once the model is written in first differences. If the ε_{it}'s are serially uncorrelated, those of the model, given by $\Delta\varepsilon_{it} = \varepsilon_{it} - \varepsilon_{it-1}$ follow an $MA(1)$ process and thus, are not correlated at order 2. On the contrary, if $\Delta\varepsilon_{it}$ appears to be correlated of order 2, one can infer that the disturbances ε_{it} exhibit some serial correlation.

Arellano and Bond (1991) show that, under the null of no serial correlation in $\Delta\varepsilon_{it}$ at order 2, the m_2 statistic is asymptotically distributed as a standard normal variate. One rejects H_0 of no serial correlation when m_2 is less than -1.64.[17]

[16] i.e., one discards, for each individual, the first two observations of the residual vector.

[17] The test is one-sided as it is indeed unlikely that the distrurbances of the first differenced model will exhibit positive serial correlation.

One can again also appeal to the Sargan/Hansen statistic to test for serial correlation in ε_{it}. First estimate the model in first differences so that the individual effects are discarded. Under the assumption of no serial correlation in ε_{it}, $\Delta\varepsilon_{it}$ follow an $MA(1)$ process and, as shown above. the series $y_{i,t-2}, y_{i,t-3}$, etc. are valid instruments for estimating this model. However, if the ε_{it}'s are serially correlated, this series no longer constitutes a valid instrument set.

This implies that one can test H_0 (ε_{it} is serially uncorrelated) against H_1 by comparing the difference between Sargan/Hansen statistics corresponding to two instrument sets: Z_0 which contains the instruments defined by the series $y_{i,t-2}, y_{i,t-3}, \ldots$ and Z_1, where Z_1 is an instrument set not dependent on the assumption of ε_{it} not being serially correlated. Indeed, to increase the test's power, one might be more specific for H_1 and test H_0 against H_1 with the latter hypothesizing, for example, that the ε_{it}'s follow an $MA(1)$ process. In this case, one would compare the Sargan/Hansen statistics associated with Z_0 as above, but with Z_1 comprising of $y_{i,t-3}$ as one of the instruments (since this would be a valid choice even under H_1). Denote the difference between the two Sargan/Hansen statistics by DQ_{sh}. Under the null this is distributed as $\chi^2_{p_0-p_1}$ where p_0 is the number of instruments in Z_0 and p_1 that in Z_1.

8.6.4 Significance Testing in Two-Step Variants

Arguably the most frequently used approach in estimating dynamic panel data models, is to follow the approach of Arellano and Bond (1991), and in particular the two-step GMM variant (popularity of these approaches may be due to the fact that GAUSS code has long been made available by the authors, and now estimation can be undertaken routinely in STATA and LIMDEP, for example).[18] A major drawback to this approach though, is that the two-step standard errors have been shown to be unreliable (Arellano and Bond, 1991; Blundell and Bond, 1998; Harris and Mátyás, 2004 and others), so that often researchers base inference on the two-step parameter estimates but using standard errors obtained in the first step. Essentially, the problem arises as the standard expressions for the conventional asymptotic variances omit the extra variation in the efficient GMM weighting matrix, $V_N(\widehat{\gamma}_1)$, which is based on the one-step estimates, $\widehat{\gamma}_1$ where

$$V_N(\widehat{\gamma}_1) = \frac{1}{N}\sum_{i=1}^{N} Z_i'\widehat{\Delta\varepsilon}_{1i}\widehat{\Delta\varepsilon}_{1i}'Z_i$$

and where $\widehat{\Delta\varepsilon}_{1i}$ are the residuals from the first-step GMM estimator.

The formula for the small sample bias corrected variance, $\widehat{V}_{bc}(\widehat{\gamma}_2)$, of $\widehat{\gamma}_2$ is rather complicated and is given by Windmeijer (2005) as

[18] There has also been a recent rise in the number of applications using the Blundell and Bond (1998) estimator.

$$\begin{aligned}\widehat{V}_{bc}(\widehat{\gamma}_2) = N\left[\Delta X'ZV_N^{-1}(\widehat{\gamma}_1)Z'\Delta X\right]^{-1} \\ + ND\left[\Delta X'ZV_N^{-1}(\widehat{\gamma}_1)Z'\Delta X\right]^{-1} \\ + N\left[\Delta X'ZV_N^{-1}(\widehat{\gamma}_1)Z'\Delta X\right]^{-1}D' \\ + D\widehat{V}(\widehat{\gamma}_1)D'\end{aligned}$$

where, using the notation $D[.,j]$ to denote all rows of the jth column of the matrix D

$$\begin{aligned}D[.,j] = -\left(\Delta X'ZV_N(\widehat{\gamma}_1)Z'\Delta X\right)^{-1}\Delta X'ZV_N(\widehat{\gamma}_1)\times \\ \left[-\frac{1}{N}\sum_{i=1}^{N}Z_i'\left(\Delta x_{ij}\Delta\widehat{\varepsilon}_{1i}' + \Delta\widehat{\varepsilon}_{1i}\Delta x_{ij}'\right)Z_i\right]V_N^{-1}(\widehat{\gamma}_1)Z'\Delta\widehat{\varepsilon}_2\end{aligned}$$

and x_{ij} is the $(T\times 1)$ jth column of X_i. Monte Carlo experiments suggest this correction does indeed work well (Windmeijer, 2005). This approach of testing linear restrictions in dynamic linear panel data models estimated using one- and two-step GMM using linear moment conditions, is further developed in Bond and Windmeijer (2005), who consider bootstrapped versions of this *Wald* test and the *LM* test (as well as three other criterion-based tests).

References

Ahn S.C. and P. Schmidt (1995), Efficient Estimation of Models for Dynamic Panel Data, *Journal of Econometrics*, Vol. 68, No. 1, pp. 5–27.

Ahn S.C. and P. Schmidt (1997), Efficient Estimation of Dynamic Panel Data Models: Alternative Assumptions and Simplified Estimation, *Journal of Econometrics*, Vol. 76, No. 1–2, pp. 309–321.

Ahn S.C. and P. Schmidt (1999), Estimation of Linear Panel Data Models Using GMM, in *Generalised Method of Moments Estimation*, L. Mátyás ed., pp. 211–245, Cambridge University Press, Cambridge.

Alvarez J. and M. Arellano (2003), The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators, *Econometrica*, Vol. 71, pp. 1121–1159.

Amemiya T. and T.E. MaCurdy (1986), Instrumental Estimation of an Error Components Model, *Econometrica*, Vol. 54, pp. 869–881.

Anderson T.W. and C. Hsiao (1982), Formulation and Estimation of Dynamic Models Using Panel Data, *Journal of Econometrics*, Vol. 18, pp. 578–606.

Arellano, M. (1989), A Note on the Anderson–Hsiao Estimator for Panel Data, *Economics Letters*, Vol. 31, pp. 337–341.

Arellano M. and S. Bond (1991), Some Tests of Specification for Panel Data: Monte-Carlo Evidence and an Application to Employment Equations, *Review of Economic Studies*, Vol. 58, pp. 127–134.

Arellano M. and O. Bover (1995), Another Look at the Instrumental Variables Estimation of Error-Components Models, *Journal of Econometrics*, Vol. 68, No. 1, pp. 29–51.

Balestra P. and M. Nerlove (1966), Pooling Cross-Section and Time-Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas, *Econometrica*, Vol. 34, pp. 585–612.

Bhargava, A. and J.D. Sargan (1983), Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods, *Econometrica*, Vol. 51, pp. 1635–1659.

Blundell R. and S. Bond (1998), Initial Conditions and Moment Restrictions in Dynamic Panel Data Models, *Journal of Econometrics*, Vol. 87, pp. 115–143.

Blundell R. and R. Smith (1991), Conditions initiales et estimation efficace dans les modèles dynamiques sur données de panel, *Annales d'Economie et de Statistique*, No. 20–21, pp. 109–123.

Bond S. and F. Windmeijer (2005), Finite Sample Inference for GMM Estimators in Linear Panel Data Models, *Econometric Reviews*, Vol. 24, pp. 1–37.

Bound J., D.A. Jaeger and R.M. Baker (1995), Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak, *Journal of the American Statistical Association*, Vol. 90, pp. 443–450.

Breusch T.S., Mizon G.E. and P. Schmidt (1989), Efficient Estimation Using Panel Data, *Econometrica*, Vol. 57, pp. 695–700.

Chamberlain, G. (1982), Multivariate Regression Models for Panel Data, *Journal of Econometrics*, Vol. 18, No. 1, pp. 5–46.

Hahn J. and J. Hausman (2002), A New Specification Test for the Validity of Instrumental Variables, *Econometrica*, Vol. 70, pp. 163–189.

Hansen L. (1982), Large Sample Properties of Generalised Method of Moments Estimators, *Econometrica*, Vol. 50, pp. 1029–1054.

Harris M.N. and L. Mátyás (1996), The Robustness of Estimators for Dynamic Panel Data Models to Misspecification, Monash University Working Paper, No. 9/96.

Harris M.N. and L. Mátyás (2000), Performance of the Operational Wansbeek–Bekker Estimator for Dynamic Panel Data Models, *Applied Economics Letters*, Vol. 7, No. 3, pp. 149–153.

Harris M.N. and L. Mátyás (2004), A Comparative Analysis of Different IV and GMM Estimators of Dynamic Panel Data Models, *International Statistical Review*, Vol. 72, No. 3, pp. 397–408.

Hausman J. (1978), Specification Tests in Econometrics, *Econometrica*, Vol. 46, pp. 1251–1271.

Hausman J. and W.E. Taylor (1981), Panel Data and Unobservable Individual Effects, *Econometrica*, Vol. 49, pp. 1377–1398.

Judson R.A. and A.L. Owen, 1999, Estimating Dynamic Panel Data Models: A guide for Macroeconomists, *Economics Letters*, Vol. 65, pp. 9–15.

Kinal T.W. (1980), The Existence of K-Class Estimators, *Econometrica*, Vol. 48, pp. 241–249.

Kiviet J.F. (1995), On Bias, Inconsistency and Efficiency of Various Estimators in Dynamic Panel Data Models, *Journal of Econometrics*, Vol. 68, No. 1, pp. 53–78.

Lee M., R. Longmire, L. Mátyás and M.N. Harris (1998), Growth Convergence: Some Panel Data Evidence, *Applied Economics*, Vol. 30, No. 7, pp. 907–912.

Maddala G.S. (1971), The Use of Variance Components Models in Pooling Cross-Section and Time Series Data, *Econometrica*, Vol. 39, pp. 341–358.

Mairesse J., B. Hall, B. Mulkay (1999), Firm-Level Investment in France and the United States: An Exploration of What We Have Learned in Twenty Years, *Annales d'Economie et de Statistique*, No. 55–56, pp. 27–68.

Nerlove M. (2002), *Essays in Panel Data Econometrics*, Cambridge University Press, Cambridge.

Nickell S. (1981), Biases in Models with Fixed Effects, *Econometrica*, Vol. 49, pp. 1417–1426.

Poskitt D.S. and C.L. Skeels (2004), Assessing the Magnitude of the Concentration Parameter in a Simultaneous Equations Model, Department of Econometrics and Business Statistcs, Monash University, Working Paper, No. 29/04.

Sargan J.D. (1958), The Estimation of Economic Relationships Using Instrumental Variables, *Econometrica*, Vol. 26, pp. 393–415.

Sevestre P. (1984), Modèles dynamiques à erreurs composées, Ph.D. Dissertation, Université Paris 1, Panthéon-Sorbonne.

Sevestre P. and A. Trognon (1985), A Note on Autoregressive Error Component Models, *Journal of Econometrics*, Vol. 28, No. 2, pp. 231–245.

Sevestre P. and A. Trognon (1983), Proprietes de Grands Echantillons d'une Classe d'Estimateurs Des Modèles Autorégressifs à Erreurs Composées, *Annales de l'INSEE*, No. 50, pp. 25–49.

Sevestre P. and Trognon, A. (1990), Consistent Estimation Methods for Dynamic Error Component Models: Small and Large Sample Properties, ERUDITE WP No. 90–03.

Shea J. (1997), Instrumental Relevance in Multivariate Linear Models: A Simple Measure, *The Review of Economics and Statistics*, Vol. 79, pp. 348–352.

Trognon, A. (1978), Miscellaneous Asymptotic Properties of Ordinary Least Squares and Maximum-Likelihood Methods in Dynamic Error-Components Models, *Annales de l'INSEE*, Vol. 30–31, pp. 631–657.

Wang J. and E. Zivot (1998), Inference on Structural Parameters in Instrumental Variables Regression with Weak Instruments, *Econometrica*, Vol. 66, pp. 1389–1404.

Wansbeek T. and P. Bekker (1996), On IV, GMM and ML in a Dynamic Panel Data Model, *Economic Letters*, Vol. 51, No. 2, pp. 145–152.

Windmeijer F. (2005), A Finite Sample Correction for the Variance of Linear Efficient Two-Step GMM Estimators, *Journal of Econometrics*,Vol. 126, No. 1, pp. 25–51.

Woglom G. (2001), More Results on the Exact Small Sample Properties of the Instrumental Variable Estimator, *Econometrica*, Vol. 69, pp. 1381–1389.

Chapter 9
Unit Roots and Cointegration in Panels

Jörg Breitung and M. Hashem Pesaran

9.1 Introduction

Recent advances in time series econometrics and panel data analysis have focussed attention on unit root and cointegration properties of variables observed over a relatively long span of time across a large number of cross section units, such as countries, regions, companies or even households. Such panel data sets have been used predominately in testing the purchasing power parity and output convergence, although the panel techniques have also been adapted more recently to the analysis of business cycle synchronization, house price convergence, regional migration and household income dynamics. This paper provides a review of the theoretical literature on testing for unit roots and cointegration in panels where the time dimension (T), and the cross section dimension (N) are relatively large. In cases where N is large (say over 100) and T small (less than 10) the analysis can proceed only under restrictive assumptions such as dynamic homogeneity and/or local cross section dependence as in spatial autoregressive or moving average models. In cases where N is small (less than 10) and T is relatively large standard time series techniques applied to systems of equations, such as the Seemingly Unrelated Regression Equations (SURE), can be used and the panel aspect of the data should not pose new technical difficulties.

One of the primary reasons behind the application of unit root and cointegration tests to a panel of cross section units was to gain statistical power and to improve on the poor power of their univariate counterparts. This was supported by the application of what might be called the first generation panel unit root tests to

Jörg Breitung
University of Bonn, Institute of Econometrics, Adenauerallee 24-42, 53113 Bonn, Germany,
e-mail: breitung@uni-bonn.de

M. Hashem Pesaran
Cambridge University and USC, Sidgwick Avenue, Cambridge, CB3 9DD, United Kingdom,
e-mail: mhp1@econ.cam.ac.uk

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

real exchange rates, output and inflation. For example, the augmented Dickey and Fuller (1979) test is typically not able to reject the hypothesis that the real exchange rate is nonstationary. In contrast, panel unit root tests applied to a collection of industrialized countries generally find that real exchange rates are stationary, thereby lending empirical support to the purchasing power parity hypothesis (e.g. Coakley and Fuertes 1997 and Choi 2001).

Unfortunately, testing the unit root and cointegration hypotheses by using panel data instead of individual time series involves several additional complications. First, panel data generally introduce a substantial amount of unobserved heterogeneity, rendering the parameters of the model cross section specific. Second, in many empirical applications, particularly the application to the real exchange rates mentioned above, it is inappropriate to assume that the cross section units are independent. To overcome these difficulties, variants of panel unit root tests are developed that allow for different forms of cross sectional dependence.[1] Third, the panel test outcomes are often difficult to interpret if the null of the unit root or cointegration is rejected. The best that can be concluded is that "a significant fraction of the cross section units is stationary or cointegrated". The panel tests do not provide explicit guidance as to the size of this fraction or the identity of the cross section units that are stationary or cointegrated. Fourth, with unobserved $I(1)$ (i.e. integrated of order unity) common factors affecting some or all the variables in the panel, it is also necessary to consider the possibility of cointegration between the variables across the groups (cross section cointegration) as well as within group cointegration. Finally, the asymptotic theory is considerably more complicated due to the fact that the sampling design involves a time as well as a cross section dimension. For example, applying the usual Dickey–Fuller test to a panel data set introduces a bias that is not present in the case of a univariate test. Furthermore, a proper limit theory has to take into account the relationship between the increasing number of time periods and cross section units (cf. Phillips and Moon 1999).

By comparison to panel unit root tests, the analysis of cointegration in panels is still at an early stage of it's developments. So far the focus of the panel cointegration literature has been on residual based approaches, although there has been a number of attempts at the development of system approaches as well. As in the case of panel unit root tests, such tests are developed based on homogenous and heterogeneous alternatives. The residual based tests were developed to ward against the "spurious regression" problem that can also arise in panels when dealing with $I(1)$ variables. Such tests are appropriate when it is known a priori that at most there can be only one within group cointegration in the panel. System approaches are required in more general settings where more than one within group cointegrating relation might be present, and/or there exist unobserved common $I(1)$ factors.

Having established a cointegration relationship, the long-run parameters can be estimated efficiently using techniques similar to the ones proposed in the case of single time series models. Specifically, fully-modified OLS procedures, the

[1] In fact the application of the second generation panel unit root tests to real exchange rates tend to over-turn the earlier test results that assume the cross section units are independently distributed. See Moon and Perron (2004) and Pesaran (2008).

dynamic OLS estimator and estimators based on a vector error correction representation were adopted to panel data structures. Most approaches employ a homogenous framework, that is, the cointegration vectors are assumed to be identical for all panel units, whereas the short-run parameters are panel specific. Although such an assumption seems plausible for some economic relationships (like the PPP hypothesis mentioned above) there are other behavioral relationships (like the consumption function or money demand), where a homogeneous framework seems overly restrictive. On the other hand, allowing all parameters to be individual specific would substantially reduce the appeal of a panel data study. It is therefore important to identify parameters that are likely to be similar across panel units whilst at the same time allowing for sufficient heterogeneity of other parameters. This requires the development of appropriate techniques for testing the homogeneity of a sub-set of parameters across the cross section units. When N is small relative to T, standard likelihood ratio based statistics can be used. Groen and Kleibergen (2003) provide an application. Testing for parameter homogeneity in the case of large panels poses new challenges that require further research. Some initial attempts are made in Pesaran, Smith and Im (1996), Phillips and Sul (2003a) and Pesaran and Yamagata (2008).

This paper reviews some recent work in this rapidly developing research area and thereby updating the earlier excellent surveys of Banerjee (1999), Baltagi and Kao (2000) and Choi (2006). The remainder of the paper is organized as follows: Sect. 9.2 sets out the basic model for the panel unit root tests and describes the first generation panel unit root tests. Second generation panel unit root tests are described in Sect. 9.3, and a brief account of the small sample properties of the panel unit root tests is provided in Sect. 9.5. General issues surrounding panel cointegration, including the problem of cross-section cointegration, are discussed in Sect. 9.6. Residual-based and system approaches to testing for cointegration in panels are reviewed in Sect. 9.7 and 9.8; and estimation of the cointegration relations in panels is discussed in Sect. 9.9. Panels with unobserved common factors, allowing for cross-section cointegration, are reviewed in Sect. 9.10. Some concluding remarks are provided in Sect. 9.11.

9.2 First Generation Panel Unit Root Tests

9.2.1 The Basic Model

Assume that time series $\{y_{i0}, \ldots, y_{iT}\}$ on the cross section units $i = 1, 2, \ldots, N$ are generated for each i by a simple first-order autoregressive, AR(1), process

$$y_{it} = (1 - \alpha_i)\mu_i + \alpha_i y_{i,t-1} + \varepsilon_{it} \,, \tag{9.1}$$

where the initial values, y_{i0}, are given, and the errors ε_{it} are identically, independently distributed (i.i.d.) across i and t with $E(\varepsilon_{it}) = 0$, $E(\varepsilon_{it}^2) = \sigma_i^2 < \infty$ and

$E(\varepsilon_{it}^4) < \infty$. These processes can also be written equivalently as simple Dickey–Fuller (DF) regressions

$$\Delta y_{it} = -\phi_i \mu_i + \phi_i y_{i,t-1} + \varepsilon_{it} \ , \tag{9.2}$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$, $\phi_i = \alpha_i - 1$. In further developments of the model it is also helpful to write (9.1) or (9.2) in mean-deviations forms $\tilde{y}_{it} = \alpha_i \tilde{y}_{i,t-1} + \varepsilon_{it}$, where $\tilde{y}_{it} = y_{it} - \mu_i$. The corresponding DF regression in $\tilde{y}_{it}$ is given by

$$\Delta \tilde{y}_{it} = \phi_i \tilde{y}_{i,t-1} + \varepsilon_{it} \ . \tag{9.3}$$

The null hypothesis of interest is

$$H_0 : \ \phi_1 = \cdots = \phi_N = 0 \ , \tag{9.4}$$

that is, all time series are independent random walks. We will consider two alternatives:

$$H_{1a} : \ \phi_1 = \cdots = \phi_N \equiv \phi \text{ and } \phi < 0$$
$$H_{1b} : \ \phi_1 < 0 \ , \ \cdots \ , \ \phi_{N_0} < 0, \ N_0 \leq N \ .$$

Under H_{1a} it is assumed that the autoregressive parameter is identical for all cross section units (see, for example, Levin and Lin (1993, LL), and Levin, Lin and Chu, 2002). This is called the *homogeneous alternative*. H_{1b} assumes that N_0 of the N $(0 < N_0 \leq N)$ panel units are stationary with individual specific autoregressive coefficients. This is referred to as the heterogeneous alternatives (see, for example, Im, Pesaran and Shin (2003, IPS). For the consistency of the test it is assumed that $N_0/N \to \kappa > 0$ as $N \to \infty$. Different panel testing procedures can be developed depending on which of the two alternatives is being considered. The panel unit root statistics motivated by the first alternative, H_{1a}, pools the observations across the different cross section units before forming the "pooled" statistic, whilst the tests developed against the heterogeneous alternatives, H_{1b}, operates directly on the test statistics for the individual cross section units using (standardized) simple averages of the underlying individual statistics or their suitable transformations such as rejection probabilities. Despite the differences in the way the two tests view the alternative hypothesis both tests can be consistent against both types of the alternatives. Also interpretation of the outcomes of both tests is subject to similar considerations discussed in the introduction. When the null hypothesis is rejected one can only conclude that a significant fraction of the AR(1) processes in the panel does not contain unit roots.

9.2.2 Derivation of the Tests

The various first generation panel unit roots proposed in the literature can be obtained using the pooled log-likelihood function of the individual Dickey–Fuller regressions given by (9.2).

$$\ell_{NT}(\phi,\theta)=\sum_{i=1}^{N}\left\{-\frac{T}{2}\log 2\pi\sigma_i^2-\frac{1}{2\sigma_i^2}\sum_{t=1}^{T}\left(\Delta y_{it}+\phi_i\mu_i-\phi_i y_{i,t-1}\right)^2\right\}, \tag{9.5}$$

where $\phi=(\phi_1,\ldots,\phi_N)'$, $\theta_i=(\mu_i,\sigma_i^2)'$ and $\theta=(\theta_1',\ldots,\theta_N')'$. In the case of the homogeneous alternatives, H_{1a}, where $\phi_i=\phi$, the maximum likelihood estimator of ϕ is given by

$$\hat{\phi}(\theta)=\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_i^{-2}\Delta y_{it}\left(y_{i,t-1}-\mu_i\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma_i^{-2}\left(y_{i,t-1}-\mu_i\right)^2}. \tag{9.6}$$

The nuisance cross-section specific parameters θ_i can be estimated either under the null or the alternative hypothesis. Under the null hypothesis μ_i is unidentified, but as we shall see it is often replaced by y_{i0}, on the implicit (identifying) assumption that $\tilde{y}_{i0}=0$ for all i. For this choice of μ_i the effective number of time periods used for estimation of ϕ_i is reduced by one. Under the alternative hypothesis the particular estimates of μ_i and σ_i^2 chosen naturally depend on the nature of the alternatives envisaged. Under homogeneous alternatives, $\phi_i=\phi<0$, the ML estimates of μ_i and σ_i^2 are given as non-linear functions of $\hat{\phi}$. Under heterogeneous alternatives ϕ_i and σ_i^2 can be treated as free parameters and estimated separately for each i.

Levin et al. (2002) avoid the problems associated with the choice of the estimators for μ_i and base their tests on the t-ratio of ϕ in the pooled fixed effect regression

$$\Delta y_{it}=a_i+\phi y_{i,t-1}+\varepsilon_{it},\ \varepsilon_{it}\backsim i.i.d.(0,\sigma_i^2).$$

The t-ratio of the FE estimator of ϕ is given by

$$\tau_{\phi}=\frac{\sum_{i=1}^{N}\widehat{\sigma}_i^{-2}\Delta\mathbf{y}_i'\mathbf{M}_e\mathbf{y}_{i,-1}}{\sqrt{\sum_{i=1}^{N}\widehat{\sigma}_i^{-2}\left(\mathbf{y}_{i,-1}'\mathbf{M}_e\mathbf{y}_{i,-1}\right)}} \tag{9.7}$$

where $\Delta\mathbf{y}_i=(\Delta y_{i1},\Delta y_{i2},\ldots,\Delta y_{iT})'$, $\mathbf{y}_{i,-1}=(y_{i0},y_{i1},\ldots,y_{i,T-1})'$, $\mathbf{M}_e=\mathbf{I}_T-\mathbf{e}_T(\mathbf{e}_T'\mathbf{e}_T)^{-1}\mathbf{e}_T'$, $\mathbf{e}_T$ is a $T\times 1$ vector of ones,

$$\hat{\sigma}_i^2=\frac{\Delta\mathbf{y}_i'\mathbf{M}_i\Delta\mathbf{y}_i}{T-2}, \tag{9.8}$$

$\mathbf{M}_i=\mathbf{I}_T-\mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'$, and $\mathbf{X}_i=(\mathbf{e}_T,\mathbf{y}_{i,-1})$.

The construction of a test against H_{1b} is less clear because the alternative consists of a set of inequality conditions. Im, Pesaran and Shin (1995, 2003) suggest the mean of the individual specific t-statistics[2]

[2] Andrews (1998) has considered optimal tests in such situations. His directed Wald statistic that gives high weights to alternatives close to the null (i.e. the parameter c in Andrews (1998) tends to zero) is equivalent to the mean of the individual specific test statistics.

$$\bar{\tau} = \frac{1}{N}\sum_{i=1}^{N} \tau_i \,, \tag{9.9}$$

where

$$\tau_i = \frac{\Delta \mathbf{y}_i' \mathbf{M}_e \mathbf{y}_{i,-1}}{\hat{\sigma}_i \left(\mathbf{y}_{i,-1}' \mathbf{M}_e \mathbf{y}_{i,-1} \right)^{1/2}} \,, \tag{9.10}$$

is the Dickey–Fuller t-statistic of cross section unit i.[3] LM versions of the t-ratios of ϕ and ϕ_i, that are analytically more tractable, can also be used which are given by

$$\tilde{\tau}_\phi = \frac{\sum_{i=1}^{N} \tilde{\sigma}_i^{-2} \Delta \mathbf{y}_i' \mathbf{M}_e \mathbf{y}_{i,-1}}{\sqrt{\sum_{i=1}^{N} \tilde{\sigma}_i^{-2} \left(\mathbf{y}_{i,-1}' \mathbf{M}_e \mathbf{y}_{i,-1} \right)}} \,, \tag{9.11}$$

and

$$\tilde{\tau}_i = \frac{\Delta \mathbf{y}_i' \mathbf{M}_e \mathbf{y}_{i,-1}}{\tilde{\sigma}_i \left(\mathbf{y}_{i,-1}' \mathbf{M}_e \mathbf{y}_{i,-1} \right)^{1/2}} \,, \tag{9.12}$$

where $\tilde{\sigma}_i^2 = (T-1)^{-1} \Delta \mathbf{y}_i' \mathbf{M}_e \Delta \mathbf{y}_i$. It is easily established that the panel unit root tests based on τ_ϕ and $\tilde{\tau}_\phi$ in the case of the pooled versions, and those based on $\bar{\tau}$ and

$$\bar{\tilde{\tau}} = N^{-1} \sum_{i=1}^{N} \tilde{\tau}_i \tag{9.13}$$

in the case of their mean group versions are asymptotically equivalent.

9.2.3 Null Distribution of the Tests

To establish the distribution of $\tilde{\tau}_\phi$ and $\bar{\tilde{\tau}}$, we first note that under $\phi_i = 0$, $\Delta \mathbf{y}_i = \sigma_i \mathbf{v}_i = \sigma_i (v_{i1}, v_{i2}, \ldots, v_{iT})'$, where $\mathbf{v}_i \backsim (\mathbf{0}, \mathbf{I}_T)$ and $\mathbf{y}_{i,-1}$ can be written as

$$\mathbf{y}_{i,-1} = y_{i0} \mathbf{e}_T + \sigma_i \mathbf{s}_{i,-1} \,, \tag{9.14}$$

where y_{i0} is a given initial value (fixed or random), $\mathbf{s}_{i,-1} = (s_{i0}, s_{i1}, \ldots, s_{i,T-1})'$, with $s_{it} = \sum_{j=1}^{t} v_{ij}$, $t = 1, 2, \ldots, T$, and $s_{i0} = 0$. Using these results in (9.11) and (9.12) we have

[3] The mean of other unit-root test statistics may be used as well. For example, Smith, Leybourne, Kim and Newbold (2004) suggest to use the mean of the weighted symmetric test statistic proposed for single time series by Park and Fuller (1995) and Fuller (1996, Sect. 10.1.3), or the Max-ADF test proposed by Leybourne (1995) based on the maximum of the original and the time reversed Dickey–Fuller test statistics.

$$\tilde{\tau}_{\phi} = \frac{\sum_{i=1}^{N}\left(\frac{\sqrt{T-1}\mathbf{v}_i'\mathbf{M}_e\mathbf{s}_{i,-1}}{\mathbf{v}_i'\mathbf{M}_e\mathbf{v}_i}\right)}{\sqrt{\sum_{i=1}^{N}\left(\frac{\mathbf{s}_{i,-1}'\mathbf{M}_e\mathbf{s}_{i,-1}}{\mathbf{v}_i'\mathbf{M}_e\mathbf{v}_i}\right)}}, \tag{9.15}$$

and

$$\bar{\tilde{\tau}} = N^{-1}\sum_{i=1}^{N}\frac{\sqrt{T-1}\mathbf{v}_i'\mathbf{M}_e\mathbf{s}_{i,-1}}{\left(\mathbf{v}_i'\mathbf{M}_e\mathbf{v}_i\right)^{1/2}\left(\mathbf{s}_{i,-1}'\mathbf{M}_e\mathbf{s}_{i,-1}\right)^{1/2}}. \tag{9.16}$$

It is clear that under the null hypothesis both test statistics are free of nuisance parameters and their critical values can be tabulated for all combinations of N and T assuming, for example, that ε_{it} (or v_{it}) are normally distributed. Therefore, in the case where the errors, ε_{it}, are serially uncorrelated an exact sample panel unit root test can be developed using either of the test statistics and no adjustments to the test statistics are needed. The main difference between the two tests lies in the way information on individual units are combined and their relative small sample performance would naturally depend on the nature of the alternative hypothesis being considered.

Asymptotic null distributions of the tests can also be derived depending on whether $(T,N) \to \infty$, sequentially, or when both N and $T \to \infty$, jointly. To derive the asymptotic distributions we need to work with the standardized versions of the test statistics

$$Z_{\text{LL}} = \frac{\tau_{\phi} - E\left(\tau_{\phi}\right)}{\sqrt{\text{Var}\left(\tau_{\phi}\right)}}, \tag{9.17}$$

and

$$Z_{\text{IPS}} = \frac{\sqrt{N}\left[\bar{\tau} - E(\tau_i)\right]}{\sqrt{\text{Var}(\tau_i)}}, \tag{9.18}$$

assuming that T is sufficiently large such that the second order moments of τ_i and τ_{ϕ} exist. The conditions under which τ_i has a second order moment are discussed in IPS and it is shown that when the underlying errors are normally distributed the second order moments exist for $T > 5$. For non-normal distributions the existence of the moments can be ensured by basing the IPS test on suitably truncated versions of the individual t-ratios. (see Pesaran (2008) for further details). The exact first and second order moments of τ_i and $\tilde{\tau}_i$ for different values of T are given in IPS (2003, Table 1). Using these results it is also possible to generalize the IPS test for unbalanced panels. Suppose the number of time periods available on the ith cross section unit is T_i, the standardized IPS statistics will now be given by

$$Z_{\text{IPS}} = \frac{\sqrt{N}\left[\bar{\tau} - N^{-1}\sum_{i=1}^{N}E(\tau_{iT_i})\right]}{\sqrt{N^{-1}\sum_{i=1}^{N}\text{Var}(\tau_{iT_i})}}, \tag{9.19}$$

where $E(\tau_{iT_i})$ and $\text{Var}(\tau_{iT_i})$ are, respectively, the exact mean and variance of the DF statistics based on T_i observations. IPS show that for all finite $T_i > 6$, $Z_{\text{IPS}} \xrightarrow{d} \mathcal{N}(0,1)$ as $N \to \infty$. Similar results follow for the LL test.

To establish the asymptotic distribution of the panel unit root tests in the case of $T \to \infty$, we first note that for each i

$$\tau_i \xrightarrow{d} \eta_i = \frac{\int_0^1 \widetilde{W}_i(a)\mathrm{d}\widetilde{W}_i(a)}{\int_0^1 \widetilde{W}_i(a)^2\mathrm{d}a} ,$$

where $\widetilde{W}_i(a)$ is a demeaned Brownian motion defined as $\widetilde{W}_i(a) = W_i(a) - \int_0^1 W_i(a)\mathrm{d}a$ and $W_1(a), \ldots, W_N(a)$ are independent standard Brownian motions. The existence of the moments of η_i are established in Nabeya (1999) who also provides numerical values for the first six moments of the DF distribution for the three standard specifications; namely models with and without intercepts and linear trends. Therefore, since the individual Dickey–Fuller statistics $\tau_1, \ldots, \tau_N$ are independent, it follows that $\eta_1, \eta_2, \ldots \eta_N$ are also independent with finite moments. Hence, by standard central limit theorems we have

$$Z_{\text{IPS}} \xrightarrow[T\to\infty]{d} \frac{\sqrt{N}[\bar{\eta} - E(\eta_i)]}{\sqrt{\text{Var}(\eta_i)}} \xrightarrow[N\to\infty]{d} \mathcal{N}(0,1) ,$$

where $\bar{\eta} = N^{-1}\sum_{i=1}^N \eta_i$. Similarly,

$$Z_{\text{LL}} = \frac{\tau_\phi - E(\tau_\phi)}{\sqrt{\text{Var}(\tau_\phi)}} \xrightarrow[(T,N)\to\infty]{d} \mathcal{N}(0,1) .$$

To simplify the exposition the above asymptotic results are derived using a sequential limit theory, where $T \to \infty$ is followed by $N \to \infty$. However, Phillips and Moon (1999) show that sequential convergence does not imply joint convergence so that in some situations the sequential limit theory may break down. In the case of models with serially uncorrelated errors, IPS (2003) show that the t-bar test is in fact valid for N and $T \to \infty$ jointly. Also as we shall see it is conjectured that the IPS test is valid for the case of serially correlated errors as N and $T \to \infty$ so long as $N/T \to k$ where k is a finite non-zero constant.

Maddala and Wu (1999) and Choi (2001) independently suggested a test against the heterogenous alternative H_{1b} that is based on the p-values of the individual statistic as originally suggested by Fisher (1932). Let π_i denote the p-value of the individual specific unit-root test applied to cross-section unit i. The combined test statistic is

$$\overline{\pi} = -2\sum_{i=1}^N \log(\pi_i) . \tag{9.20}$$

Another possibility would be to use the inverse normal test defined by

$$Z_{\text{INV}} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Phi^{-1}(\pi_i) \,, \tag{9.21}$$

where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. An important advantage of this approach is that it is possible to allow for different specifications (such as different deterministic terms and lag orders) for each panel unit.

Under the null hypothesis $\overline{\pi}$ is χ^2 distributed with $2N$ degrees of freedom. For large N the transformed statistic

$$\bar{\pi}^* = -\frac{1}{\sqrt{N}} \sum_{i=1}^{N} [\log(\pi_i) + 1] \,, \tag{9.22}$$

is shown to have a standard normal limiting null distribution as $T, N \to \infty$, sequentially.

9.2.4 Asymptotic Power of the Tests

It is interesting to compare the asymptotic power of the test statistics against the sequence of local alternatives

$$H_\ell : \quad \alpha_{i,NT} = 1 - \frac{c_i}{T\sqrt{N}} \,. \tag{9.23}$$

Following Breitung (2000) and Moon, Perron and Phillips (2007) the asymptotic distribution under H_ℓ is obtained as $Z_j \xrightarrow{d} \mathcal{N}(-\bar{c}\,\theta_j, 1)$, $j =$LL, IPS, where $\bar{c} = \lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} c_i$ and

$$\theta_1 = \sqrt{E\left(\int_0^1 \widetilde{W}_i(a)^2 \mathrm{d}a\right)}, \; \theta_2 = \frac{E\left(\sqrt{\int_0^1 \widetilde{W}_i(a)^2 \mathrm{d}a}\right)}{\sqrt{\mathrm{Var}(\tau_i)}} \,.$$

It is interesting to note that the local power of both test statistics depends on the mean $\bar{c}$. Accordingly, the test statistics do not exploit the deviations from the mean value of the autoregressive parameter.

Moon et al. (2007) derive the most powerful test statistic against the local alternative (9.23). Assume that we (randomly) choose the sequence $c_1^*, \ldots, c_N^*$ instead of the unknown values $c_1, \ldots, c_N$. The point optimal test statistic is constructed using the (local-to-unity) pseudo differences

$$\Delta_{c_i^*} y_{it} = y_{it} - (1 - c_i^*/T\sqrt{N}) y_{i,t-1} \text{ for } t = 1, \ldots, T \,.$$

For the model without individual constants and homogeneous variances the point optimal test results in the statistic

$$V_{NT} = \frac{1}{\widehat{\sigma}^2}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(\Delta_{c_i^*} y_{it})^2 - (\Delta y_{it})^2\right) - \frac{1}{2}\kappa^2 ,$$

where $E(c_i^*)^2 = \kappa^2$. Under the sequence of local alternatives (9.23) Moon et al. (2007, Theorem 7) derive the limiting distribution as

$$V_{NT} \xrightarrow{d} \mathcal{N}\left(-E(c_i c_i^*), 2\kappa^2\right) .$$

The upper bound of the local power is achieved with $c_i = c_i^*$, that is, if the local alternatives used to construct the test coincide with the actual alternative. Unfortunately, in practice it seems extremely unlikely that one could select values of c_i^* that are perfectly correlated with the true values, c_i. If, on the other hand, the variates c_i^* are independent of c_i, then the power is smaller than the power of a test using identical values $c_i^* = c^*$ for all i. This suggests that if there is no information about variation of c_i, then a test cannot be improved by taking into account a possible heterogeneity of the alternative.

9.2.5 Heterogeneous Trends

To allow for more general mean functions we consider the model:

$$y_{it} = \delta_i' \mathbf{d}_{it} + \tilde{y}_{it} , \tag{9.24}$$

where $\mathbf{d}_{it}$ represents the deterministics and $\Delta\tilde{y}_{it} = \phi_i \tilde{y}_{i,t-1} + \varepsilon_{it}$. For the model with a constant mean we let $\mathbf{d}_{it} = 1$ and the model with individual specific time trends $\mathbf{d}_{it}$ is given by $\mathbf{d}_{it} = (1, t)'$. Furthermore, structural breaks in the mean function can be accommodated by including (possibly individual specific) dummy variables in the vector $\mathbf{d}_{it}$. The parameter vector δ_i is assumed to be unknown and has to be estimated. For the Dickey–Fuller test statistic the mean function is estimated under the alternative, that is, for the model with a time trend $\widehat{\boldsymbol{\delta}}_i' \mathbf{d}_{it}$ results from a regression of y_{it} on a constant and t $(t = 1, 2, \ldots, T)$. Alternatively, the mean function can also be estimated under the null hypothesis (cf. Schmidt and Phillips, 1992) or under a local alternative (Elliott, Rothenberg and Stock, 1996).[4]

Including deterministic terms may have an important effect on the asymptotic properties of the test. Let $\Delta\widehat{\tilde{y}}_t$ and $\widehat{\tilde{y}}_{i,t-1}$ denote estimates for $\Delta\tilde{y}_{it} = \Delta y_{it} - E(\Delta y_{it})$ and $\tilde{y}_{i,t-1} = y_{i,t-1} - E(y_{i,t-1})$. In general, running the regression

$$\Delta\widehat{\tilde{y}}_{it} = \phi\widehat{\tilde{y}}_{i,t-1} + e_{it}$$

does not render a t-statistic with a standard normal limiting distribution due to the fact that $\widehat{\tilde{y}}_{i,t-1}$ is correlated with e_{it}. For example, if $\mathbf{d}_{it}$ is an individual specific

[4] See, e.g. Choi (2002) and Harvey, Leybourne and Sakkas (2006).

constant such that $\widehat{\widetilde{y}}_{it} = y_{it} - T^{-1}(y_{i0} + \cdots + y_{i,T-1})$ we obtain under the null hypothesis

$$\lim_{T\to\infty} \frac{1}{T} E\left\{ \sum_{t=1}^{T} e_{it} \widehat{\widetilde{y}}_{i,t-1} \right\} = -\sigma_i^2/2 \, .$$

It follows that the t-statistic of $\phi = 0$ tends to $-\infty$ as N or T tends to infinity.

To correct for the bias, Levin et al. (2002) suggested using the correction terms

$$a_T(\hat{\boldsymbol{\delta}}) = E\left(\frac{1}{\sigma_i^2 T} \sum_{t=1}^{T} \Delta\widehat{\widetilde{y}}_{it} \widehat{\widetilde{y}}_{i,t-1} \right) \tag{9.25}$$

$$b_T^2(\hat{\boldsymbol{\delta}}) = \frac{\mathrm{Var}\left(T^{-1} \sum_{t=1}^{T} \Delta\widehat{\widetilde{y}}_{it} \widehat{\widetilde{y}}_{i,t-1} \right)}{\sigma_i^2 E\left(T^{-1} \sum_{t=1}^{T} \widehat{\widetilde{y}}_{i,t-1}^2 \right)} \tag{9.26}$$

where $\delta = (\hat{\boldsymbol{\delta}}_1', \hat{\boldsymbol{\delta}}_2', \ldots, \hat{\boldsymbol{\delta}}_N')'$, and $\hat{\boldsymbol{\delta}}_i$ is the estimator of the coefficients of the deterministics, $\mathbf{d}_{it}$, in the OLS regression of y_{it} on $\mathbf{d}_{it}$. The corrected, standardized statitic is given by

$$Z_{\mathrm{LL}}(\hat{\boldsymbol{\delta}}) = \frac{\left[\sum_{i=1}^{N} \sum_{t=1}^{T} \Delta\widehat{\widetilde{y}}_{it} \widehat{\widetilde{y}}_{i,t-1} / \widehat{\sigma}_i^2 \right] - NTa_T(\hat{\boldsymbol{\delta}})}{b_T(\hat{\boldsymbol{\delta}}) \sqrt{\sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\widetilde{y}}_{i,t-1}^2 / \widehat{\sigma}_i^2}} \, .$$

Levin et al. (2002) present simulated values of $a_T(\hat{\boldsymbol{\delta}})$ and $b_T(\hat{\boldsymbol{\delta}})$ for models with constants, time trends and various values of T. A problem is, however, that for unbalanced data sets no correction terms are tabulated.

Alternatively, the test statistic may be corrected such that the adjusted t-statistic

$$Z_{\mathrm{LL}}^*(\hat{\boldsymbol{\delta}}) = [Z_{\mathrm{LL}}(\hat{\boldsymbol{\delta}}) - a_T^*(\hat{\boldsymbol{\delta}})]/b_T^*(\hat{\boldsymbol{\delta}})$$

is asymptotically standard normal. Harris and Tzavalis (1999) derive the small sample values of $a_T^*(\hat{\boldsymbol{\delta}})$ and $b_T^*(\hat{\boldsymbol{\delta}})$ for T fixed and $N \to \infty$. Therefore, their test statistic can be applied for small values of T and large values of N.

An alternative approach is to avoid the bias – and hence the correction terms – by using alternative estimates of the deterministic terms. Breitung and Meyer (1994) suggest using the initial value y_{i0} as an estimator of the constant term. As argued by Schmidt and Phillips (1992), the initial value is the best estimate of the constant given the null hypothesis is true. Using this approach the regression equation for a model with a constant term becomes

$$\Delta y_{it} = \phi^*(y_{i,t-1} - y_{i0}) + v_{it} \, .$$

Under the null hypothesis, the pooled t-statistic of $H_0 : \phi^* = 0$ has a standard normal limit distribution.

For a model with a linear time trend a minimal invariant statistic is obtained by the transformation (cf. Ploberger and Phillips, 2002)

$$x_{it}^* = y_{it} - y_{i0} - \frac{t}{T}(y_{iT} - y_{i0}) \, .$$

In this transformation subtracting y_{i0} eliminates the constant and $(y_{iT} - y_{i0})/T = (\Delta y_{i1} + \cdots + \Delta y_{iT})/T$ is an estimate of the slope of the individual trend function.

To correct for the mean of Δy_{it} a Helmert transformation can be used

$$\Delta y_{it}^* = s_t \left[\Delta y_{it} - \frac{1}{T-t}(\Delta y_{i,t+1} + \cdots + \Delta y_{iT}) \right], \; t = 1, \ldots, T-1$$

where $s_t^2 = (T-t)/(T-t+1)$ (cf. Arellano, 2003, p. 17). Using these transformations the regression equation becomes

$$\Delta y_{it}^* = \phi^* x_{i,t-1}^* + v_{it} \, . \tag{9.27}$$

It is not difficult to verify that under the null hypothesis $E(\Delta y_{it}^* x_{i,t-1}^*) = 0$ and, thus, the t-statistic for $\phi^* = 0$ is asymptotically standard normally distributed (cf. Breitung, 2000).

It is important to note that including individual specific time trends substantially reduce the (local) power of the test. This was first observed by Breitung (2000) and studied more rigorously by Ploberger and Phillips (2002) and Moon et al. (2007). Specifically, the latter two papers show that a panel unit root test with incidental trends has nontrivial asymptotic power only for local alternatives with rate $T^{-1}N^{-1/4}$. A similar result is found by Moon, Perron and Phillips (2006) for the test suggested by Breitung (2000).

The test against heterogeneous alternatives H_{1b} can be easily adjusted for individual specific deterministic terms such as linear trends or seasonal dummies. This can be done by computing IPS statistics, defined by (9.18) and (9.19) for the balanced and unbalanced panels, using Dickey–Fuller t-statistics based on DF regressions including the deterministics $\delta_i' \mathbf{d}_{it}$, where $\mathbf{d}_{it} = 1$ in the case of a constant term, $\mathbf{d}_{it} = (1, t)'$ in the case of models with a linear time trend and so on. The mean and variance corrections should, however, be computed to match the nature of the deterministics. Under a general setting IPS (2003) have shown that the Z_{IPS} statistic converges in distribution to a standard normal variate as $N, T \to \infty$, jointly.

In a straightforward manner it is possible to include dummy variables in the vector $\mathbf{d}_{it}$ that accommodate structural breaks in the mean function (see, e.g., Murray and Papell, 2002; Tzavalis, 2002; Carrion-I-Sevestre, Del Barrio and Lopez-Bazo, 2005; Breitung and Candelon, 2005; Im, Lee and Tieslau, 2005).

9.2.6 Short-Run Dynamics

If it is assumed that the error in the autoregression (9.1) is a serially correlated stationary process, the short-run dynamics of the errors can be accounted for by including lagged differences

$$\Delta y_{it} = \delta_i' \mathbf{d}_{it} + \phi_i y_{i,t-1} + \gamma_{i1}\Delta y_{i,t-1} + \cdots + \gamma_{i,p_i}\Delta y_{i,t-p_i} + \varepsilon_{it} \ . \tag{9.28}$$

For example, the IPS statistics (9.18) and (9.19) developed for balanced and unbalanced panels can now be constructed using the ADF(p_i) statistics based on the above regressions. As noted in IPS (2003), small sample properties of the test can be much improved if the standardization of the IPS statistic is carried out using the simulated means and variances of $\tau_i(p_i)$, the t-ratio of ϕ_i computed based on ADF(p_i) regressions. This is likely to yield better approximations, since $E[\tau_i(p_i)]$, for example, makes use of the information contained in p_i while $E[\tau_i(0)] = E(\tau_i)$ does not. Therefore, in the serially correlated case IPS propose the following standardized t-bar statistic

$$Z_{\text{IPS}} = \frac{\sqrt{N}\left\{\bar{\tau} - \frac{1}{N}\sum_{i=1}^{N} E[\tau_i(p_i)]\right\}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} Var[\tau_i(p_i)]}} \xrightarrow[(T,N)\to\infty]{d} \mathcal{N}(0,1) \ . \tag{9.29}$$

The value of $E[\tau_i(p)]$ and $\text{Var}[\tau_i(p)]$ simulated for different values of T and p, are provided in Table 3 of IPS. These simulated moments also allow the IPS panel unit root test to be applied to unbalanced panels with serially correlated errors.

For tests against the homogenous alternatives, $\phi_1 = \cdots = \phi_N = \phi < 0$, Levin et al. (2002) suggest removing all individual specific parameters within a first step regression such that e_{it} ($v_{i,t-1}$) are the residuals from a regression of Δy_{it} ($y_{i,t-1}$) on $\Delta y_{i,t-1}, \ldots, \Delta y_{i,t-p_i}$ and d_{it}. In the second step the common parameter ϕ is estimated from a pooled regression

$$(e_{it}/\widehat{\sigma}_i) = \phi(v_{i,t-1}/\widehat{\sigma}_i) + \nu_{it} \ ,$$

where $\widehat{\sigma}_i^2$ is the estimated variance of e_{it}. Unfortunately, the first step regressions are not sufficient to remove the effect of the short-run dynamics on the null distribution of the test. Specifically,

$$\lim_{T\to\infty} E\left[\frac{1}{T-p}\sum_{t=p+1}^{T} e_{it}v_{i,t-1}/\sigma_i^2\right] = \frac{\bar{\sigma}_i}{\sigma_i}\, a_\infty(\hat{\boldsymbol{\delta}}) \ ,$$

where $\overline{\sigma}_i^2$ is the long-run variance and $a_\infty(\hat{\boldsymbol{\delta}})$ denotes the limit of the correction term given in (9.25). Levin et al. (2002) propose a nonparametric (kernel based) estimator for $\bar{\sigma}_i^2$

$$\bar{s}_i^2 = \frac{1}{T}\left[\sum_{t=1}^{T}\Delta\widehat{\bar{y}}_{it}^2 + 2\sum_{l=1}^{K}\left(\frac{K+1-l}{K+1}\right)\left(\sum_{t=l+1}^{T}\Delta\widehat{\bar{y}}_{it}\Delta\widehat{\bar{y}}_{i,t-l}\right)\right], \tag{9.30}$$

where $\Delta\widehat{\bar{y}}_{it}$ denotes the demeaned difference and K denotes the truncation lag. As noted by Phillips and Ouliaris (1990), in a time series context the estimator of the long-run variance based on differences is inappropriate since under the stationary alternative $\bar{s}_i^2 \xrightarrow{p} 0$ and, thus, using this estimator yields an inconsistent test. In contrast, in the case of panels the use of $\bar{s}_i^2$ improves the power of the test, since with $\bar{s}_i^2 \xrightarrow{p} 0$ the correction term drops out and the test statistic tends to $-\infty$.

It is possible to avoid the use of a kernel based estimator of the long-run variance by using an alternative approach suggested by Breitung and Das (2005). Under the null hypothesis we have

$$\gamma_i(L)\Delta y_{it} = \delta_i'\mathbf{d}_{it} + \varepsilon_{it}\ ,$$

where $\gamma_i(L) = 1 - \gamma_{i1}L - \cdots - \gamma_{i,p_i}L^p$ and L is the lag operator. It follows that $\widetilde{g}_t = \gamma_i(L)[y_{it} - E(y_{it})]$ is a random walk with uncorrelated increments. Therefore, the serial correlation can be removed by replacing y_{it} by the pre-whitened variable $\hat{y}_{it} = \widehat{\gamma}_i(L)y_{it}$, where $\widehat{\gamma}_i(L)$ is an estimator of the lag polynomial obtained from the least-square regression

$$\Delta y_{it} = \delta_i'\mathbf{d}_{it} + \gamma_{i1}\Delta y_{i,t-1} + \cdots + \gamma_{i,p_i}\Delta y_{i,t-p_i} + \varepsilon_{it}\ . \tag{9.31}$$

This approach may also be used for modifying the "unbiased statistic" based on the t-statistic of $\phi^* = 0$ in (9.27). The resulting t-statistic has a standard normal limiting distribution if as $T \to \infty$ is followed by $N \to \infty$.

A related approach is suggested by Westerlund (2008). He suggests to test the unit root hypothesis by running a modified ADF regression of the form

$$\Delta y_{it} = \delta_i'\mathbf{d}_{it} + \phi_i y_{i,t-1}^* + \gamma_{i1}\Delta y_{i,t-1} + \cdots + \gamma_{i,p_i}\Delta y_{i,t-p_i} + \varepsilon_{it}\ . \tag{9.32}$$

where $y_{i,t-1}^* = (\widehat{\sigma}_i/\bar{s}_i)y_{i,t-1}$ and $\bar{s}_i^2$ is a consistent estimator of the long-run variance, $\overline{\sigma}_i^2$. Westerlund (2008) recommends to use a parametric estimate of the long-run variance based on an autoregressive representation. This transformation of the lagged dependent variable eliminates the nuisance parameters in the asymptotic distribution of the ADF statistic and, therefore, the correction for the numerator of the corrected t-statistic of Levin et al. (2002) is the same as in the case without short-run dynamics.

Pedroni and Vogelsang (2005) have proposed a test statistic that avoids the specification of the short-run dynamics by using an autoregressive approximation. Their test statistic is based on the pooled variance ratio statistic

$$Z_{NT}^{w} = \frac{Tc_i(0)}{N\widehat{s}_i^2}\ ,$$

where $c_i(\ell) = T^{-1}\sum_{t=\ell+1}^{T}\widehat{\tilde{y}}_{it}\widehat{\tilde{y}}_{i,t-\ell}$, $\widehat{\tilde{y}}_{it} = y_{it} - \hat{\boldsymbol{\delta}}_i'\mathbf{d}_{it}$ and $\hat{s}_i^2$ is the untruncated Bartlett kernel estimator defined as $\hat{s}_i^2 = \sum_{\ell=-T+1}^{T+1}(1-|\ell|/T)c_i(\ell)$. As has been shown by Kiefer and Vogelsang (2002) and Breitung (2002), the limiting distribution of such "nonparametric" statistics does not depend on nuisance parameters involved by the short run dynamics of the processes. Accordingly, no adjustment for short-run dynamics is necessary.

9.2.7 Other Approaches to Panel Unit Root Testing

An important problem of combining Dickey–Fuller type statistics in a panel unit root test is that they involve a nonstandard limiting distribution. If the panel unit root statistic is based on a standard normally distributed test statistic z_i, then $N^{-1/2}\sum_{i=1}^{N} z_i$ has a standard normal limiting distribution even for a finite N. In this case no correction terms need to be tabulated to account for the mean and the variance of the test statistic.

Chang (2002) proposes a nonlinear instrumental variable (IV) approach, where the transformed variable

$$w_{i,t-1} = y_{i,t-1}e^{-c_i|y_{i,t-1}|}$$

with $c_i > 0$ is used as an instrument for estimating ϕ_i in the regression $\Delta y_{it} = \phi_i y_{i,t-1} + \varepsilon_{it}$ (which may also include deterministic terms and lagged differences). Since $w_{i,t-1}$ tends to zero as $y_{i,t-1}$ tends to $\pm\infty$ the trending behavior of the nonstationary variable $y_{i,t-1}$ is eliminated. Using the results of Chang, Park and Phillips (2001), Chang (2002) showed that the Wald test of $\phi = 0$ based on the nonlinear IV estimator possesses a standard normal limiting distribution. Another important property of the test is that the nonlinear transformation also takes account of possible contemporaneous dependence among the cross section units. Accordingly, Chang's panel unit root test is also robust against cross-section dependence.

It should be noted that $w_{i,t-1} \in [-(c_i e)^{-1}, (c_i e)^{-1}]$ with a maximum (minimum) at $y_{i,t-1} = 1/c_i$ ($y_{i,t-1} = -1/c_i$). Therefore, the choice of the parameter c_i is crucial for the properties of the test. First, the parameter should be proportional to inverse of the standard deviations of Δy_{it}. Chang notes that if the time dimension is small, the test slightly over-rejects the null and therefore she proposes to use a larger value of K to correct for the size distortions.

An alternative approach to obtain an asymptotically standard normal test statistic is to adjust the given samples in all cross-sections so that they all have sums of squares $y_{i1}^2 + \cdots + y_{ik_i}^2 = \sigma_i^2 cT^2 + h_i$, where $h_i \xrightarrow{p} 0$ as $T \to \infty$. In other words, the panel data set becomes an unbalanced panel with k_i time periods in the i'th unit. Chang and Park (2004) calls this setting the "equi-squared sum contour", whereas the traditional framework is called the "equi-sample-size contour". The nice feature of this approach is that it yields asymptotically standard normal test statistics.

An important drawback is, however, that a large number of observations may be discarded by applying this contour which may result in a severe loss of power.

Hassler, Demetrescu and Tarcolea (2006) have suggested to use the LM statistic for a fractional unit root as an asymptotically normally distributed test statistic. This test statistic is uniformly most powerful against fractional alternatives of the form $(1-L)^d y_{it} = \varepsilon_{it}$ with $d < 1$. Although usually panel unit root tests are used to decide whether the series are $I(1)$ or $I(0)$, it can be argued that fractional unit root tests also have a good (albeit not optimal) power against the $I(0)$ alternative (e.g. Robinson, 1994).

As in the time series case it is possible to test the null hypothesis that the series are stationary against the alternative that (at least some of) the series are nonstationary. The test suggested by Tanaka (1990) and Kwiatkowski, Phillips, Schmidt and Shin (1992) is designed to test the hypothesis $H_0^* : \theta_i = 0$ in the model

$$y_{it} = \boldsymbol{\delta}_i' \mathbf{d}_{it} + \theta_i r_{it} + u_{it} \ , \ \ t = 1, \ldots, T \ , \tag{9.33}$$

where Δr_{it} is white noise with unit variance and u_{it} is stationary. The cross-section specific KPSS statistic is

$$\kappa_i = \frac{1}{T^2 \bar{\sigma}_{T,i}^2} \sum_{t=1}^{T} \widehat{S}_{it}^2 \ ,$$

where $\bar{\sigma}_{T,i}^2$ denotes a consistent estimator of the long-run variance of Δy_{it} and $\widehat{S}_{it} = \sum_{\ell=1}^{t} \left(y_{i\ell} - \hat{\boldsymbol{\delta}}_i' \mathbf{d}_{i\ell} \right)$ is the partial sum of the residuals from a regression of y_{it} on the deterministic terms (a constant or a linear time trend). The individual test statistics can be combined as in the test suggested by IPS (2003) yielding

$$\bar{\kappa} = N^{-1/2} \frac{\sum_{i=1}^{N} [\kappa_i - E(\kappa_i)]}{\sqrt{\mathrm{Var}(\kappa_i)}} \ ,$$

where asymptotic values of $E(\kappa_i)$ and $\mathrm{Var}(\kappa_i)$ are derived in Hadri (2000) and values for finite T and $N \to \infty$ are presented in Hadri and Larsson (2005).

The test of Harris, Leybourne and McCabe (2004) is based on the stationarity statistic

$$Z_i(k) = \sqrt{T}\, \widehat{c}_i(k) / \widehat{\omega}_{zi}(k) \ ,$$

where $\widehat{c}_i(k)$ denotes the usual estimator of the covariance at lag k of cross section unit i and $\widehat{\omega}_{zi}^2(k)$ is an estimator of the long-run variance of $z_{it}^k = (y_{it} - \hat{\boldsymbol{\delta}}_i' \mathbf{d}_{it})(y_{i,t-k} - \hat{\boldsymbol{\delta}}_i' \mathbf{d}_{i,t-k})$. The intuition behind this test statistic is that for a stationary and ergodic time series we have $E[\widehat{c}_i(k)] \to 0$ as $k \to \infty$. Since $\widehat{\omega}_{zi}^2$ is a consistent estimator for the variance of $\widehat{c}_i(k)$ it follows that $Z_i(k)$ converges to a standard normally distributed random variable as $k \to \infty$ and $k/\sqrt{T} \to \delta < \infty$.

9.3 Second Generation Panel Unit Root Tests

9.3.1 Cross-Section Dependence

So far we have assumed that the time series $\{y_{it}\}_{t=0}^{T}$ are independent across i. However, in many macroeconomic applications using country or regional data it is found that the time series are contemporaneously correlated. Prominent examples are the analysis of purchasing power parity and output convergence.[5] The literature on modelling of cross section dependence in large panels is still developing and in what follows we provide an overview of some of the recent contributions.[6]

Cross section dependence can arise due to a variety of factors, such as omitted observed common factors, spatial spill over effects, unobserved common factors, or general residual interdependence that could remain even when all the observed and unobserved common effects are taken into account. Abstracting from common observed effects and residual serial correlation a general specification for cross sectional error dependence can be written as

$$\Delta y_{it} = -\mu_i \phi_i + \phi_i y_{i,t-1} + u_{it} \ ,$$

where

$$u_{it} = \gamma_i' \mathbf{f}_t + \xi_{it} \ , \tag{9.34}$$

or

$$\mathbf{u}_t = \mathbf{\Gamma} \boldsymbol{f}_t + \boldsymbol{\xi}_t \ , \tag{9.35}$$

$\mathbf{u}_t = (u_{1t}, u_{2t}, \ldots, u_{Nt})'$, $\mathbf{f}_t$ is an $m \times 1$ vector of serially uncorrelated unobserved common factors, and $\boldsymbol{\xi}_t = (\xi_{1t}, \xi_{2t}, \ldots, \xi_{Nt})'$ is an $N \times 1$ vector of serially uncorrelated errors with mean zero and the positive definite covariance matrix $\mathbf{\Omega}_\xi$, and $\mathbf{\Gamma}$ is an $N \times m$ matrix of factor loadings defined by $\mathbf{\Gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_N)'$.[7] Without loss of generality the covariance matrix of $\mathbf{f}_t$ is set to $\mathbf{I}_m$, and it is assumed that $\mathbf{f}_t$ and $\boldsymbol{\xi}_t$ are independently distributed. If $\gamma_1 = \cdots = \gamma_N$, then $\theta_t = \gamma' \mathbf{f}_t$ is a conventional "time effect" that can be removed by subtracting the cross section means from the data. In general it is assumed that γ_i, the factor loading for the ith cross section unit, differs across i and represents draws from a given distribution.

Under the above assumptions and conditional on γ_i, $i = 1, 2, \ldots, N$, the covariance matrix of the composite errors, $\mathbf{u}_t$, is given by $\mathbf{\Omega} = \mathbf{\Gamma}\mathbf{\Gamma}' + \mathbf{\Omega}_\xi$. It is clear that without further restrictions the matrices $\mathbf{\Gamma}$ and $\mathbf{\Omega}_\xi$ are not separately identified. The properties of $\mathbf{\Omega}$ also crucially depend on the relative eigenvalues of $\mathbf{\Gamma}\mathbf{\Gamma}'$ and $\mathbf{\Omega}_\xi$, and their limits as $N \to \infty$. Accordingly two cases of cross-section dependence can

[5] See, for example, O'Connell (1998) and Phillips and Sul (2003b). Tests for cross section independence of errors with applications to output growth equations are considered in Pesaran (2004).

[6] A survey of the second generation panel unit root tests is also provided by Hurlin and Mignon (2004).

[7] The case where $\mathbf{f}_t$ and/or ξ_{it} might be serially correlated will be considered below.

be distinguished: (i) *Weak dependence.* This cases arises if it is assumed that the eigenvalues of $\mathbf{\Omega}$ are bounded as $N \to \infty$. This assumption rules out the presence of unobserved common factors, but allows the cross section units to be, for example, spatially correlated with a finite number of "neighbors". (ii) *Strong dependence.* In this case some eigenvalues of $\mathbf{\Omega}$ are $O(N)$, which arises when there are unobserved common factors. When N is fixed as $T \to \infty$ both sources of dependence could be present. But for $N \to \infty$ (and particularly when $N > T$) it seems only sensible to consider cases where $\text{rank}(\mathbf{\Gamma}) = m \geq 1$ and $\mathbf{\Omega}_\xi$ is a diagonal matrix. A general discussion of the concepts of weak and strong cross section dependence is provided in Pesaran and Tosetti (2007) where it is shown that all spatial econometric models considered in the literature are examples of weak cross section dependence.

A simple example of panel data models with weak cross section dependence is given by

$$\begin{bmatrix} \Delta y_{1t} \\ \vdots \\ \Delta y_{Nt} \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix} + \phi \begin{bmatrix} y_{1,t-1} \\ \vdots \\ y_{N,t-1} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ \vdots \\ u_{Nt} \end{bmatrix} \tag{9.36}$$

or

$$\Delta \mathbf{y}_t = \mathbf{a} + \phi \mathbf{y}_{t-1} + \mathbf{u}_t \ , \tag{9.37}$$

where $a_i = -\phi \mu_i$ and $\Delta \mathbf{y}_t$, $\mathbf{y}_{t-1}$, $\mathbf{a}$ and $\mathbf{u}_t$ are $N \times 1$ vectors. The cross-section correlation is represented by a non-diagonal matrix

$$\mathbf{\Omega} = E(\mathbf{u}_t \mathbf{u}_t'), \text{ for all } t \ ,$$

with bounded eigenvalues. For the model without constants Breitung and Das (2005) showed that the regression t-statistic of $\phi = 0$ in (9.37) is asymptotically distributed as $\mathcal{N}(0, v_\Omega)$ where

$$v_\Omega = \lim_{N \to \infty} \frac{tr(\mathbf{\Omega}^2/N)}{(tr\mathbf{\Omega}/N)^2} \ . \tag{9.38}$$

Note that $tr(\mathbf{\Omega})$ and $tr(\mathbf{\Omega}^2)$ are $O(N)$ and, thus, v_Ω converges to a constant that can be shown to be larger than one. This explains why the test ignoring the cross-correlation of the errors has a positive size bias.

9.3.2 Tests Based on GLS Regressions

Since (9.37) can be seen as a seemingly unrelated regression system, O'Connell (1998) suggests to estimate the system by using a GLS estimator (see also Flores, Jorion, Preumont and Szarfarz, 1999). Let $\widehat{\mathbf{\Omega}} = T^{-1} \sum_{t=1}^{T} \widehat{\mathbf{u}}_t \widehat{\mathbf{u}}_t'$ denote the sample covariance matrix of the residual vector. The GLS t-statistic is given by

$$t_{\mathrm{gls}}(N) = \frac{\sum_{t=1}^{T} \Delta\widetilde{\mathbf{y}}_t' \widehat{\mathbf{\Omega}}^{-1} \widetilde{\mathbf{y}}_{t-1}}{\sqrt{\sum_{t=1}^{T} \widetilde{\mathbf{y}}_{t-1}' \widehat{\mathbf{\Omega}}^{-1} \widetilde{\mathbf{y}}_{t-1}}} ,$$

where $\widetilde{\mathbf{y}}_t$ is the vector of demeaned variables. Harvey and Bates (2003) derive the limiting distribution of $t_{\mathrm{gls}}(N)$ for a fixed N and as $T \to \infty$, and tabulate its asymptotic distribution for various values of N. Breitung and Das (2005) show that if $\widetilde{\mathbf{y}}_t = \mathbf{y}_t - \mathbf{y}_0$ is used to demean the variables and $T \to \infty$ is followed by $N \to \infty$, then the GLS t-statistic possesses a standard normal limiting distribution.

The GLS approach cannot be used if $T < N$ as in this case the estimated covariance matrix $\widehat{\mathbf{\Omega}}$ is singular. Furthermore, Monte Carlo simulations suggest that for reasonable size properties of the GLS test, T must be substantially larger than N (e.g. Breitung and Das, 2005). Maddala and Wu (1999) and Chang (2004) have suggested a bootstrap procedure that improves the size properties of the GLS test.

9.3.3 Test Statistics Based on OLS Regressions

An alternative approach based on "panel corrected standard errors" (PCSE) is considered by Jönsson (2005) and Breitung and Das (2005). In the model with weak dependence, the variance of the OLS estimator $\widehat{\phi}$ is consistently estimated by

$$\widehat{\mathrm{var}}(\widehat{\phi}) = \frac{\sum_{t=1}^{T} \widetilde{\mathbf{y}}_{t-1}' \widehat{\mathbf{\Omega}} \widetilde{\mathbf{y}}_{t-1}}{\left(\sum_{t=1}^{T} \widetilde{\mathbf{y}}_{t-1}' \widetilde{\mathbf{y}}_{t-1} \right)^2} .$$

If $T \to \infty$ is followed by $N \to \infty$ the robust t statistic $t_{\mathrm{rob}} = \widehat{\phi} / \sqrt{\widehat{\mathrm{var}}(\widehat{\phi})}$ is asymptotically standard normally distributed (Breitung and Das, 2005).

If it is assumed that the cross correlation is due to common factors, then the largest eigenvalue of the error covariance matrix, $\mathbf{\Omega}$, is $O_p(N)$ and the robust PCSE approach breaks down. Specifically, Breitung and Das (2008) showed that in this case t_{rob} is distributed as the ordinary Dickey-Fuller test applied to the first principal component.

In the case of a single unobserved common factor, Pesaran (2008) has suggested a simple modification of the usual test procedure. Let $\bar{y}_t = N^{-1} \sum_{i=1}^{N} y_{it}$ and $\Delta\bar{y}_t = N^{-1} \sum_{i=1}^{N} \Delta y_{it} = \bar{y}_t - \bar{y}_{t-1}$. The cross section augmented Dickey–Fuller (CADF) test is based on the following regression

$$\Delta y_{it} = a_i + \phi_i y_{i,t-1} + b_i \bar{y}_{t-1} + c_i \Delta\bar{y}_t + e_{it} .$$

In this regression the additional variables $\Delta\bar{y}_t$ and $\bar{y}_{t-1}$ are $\sqrt{N}$-consistent estimators for the rescaled factors $\bar{\gamma}f_t$ and $\bar{\gamma}\sum_{j=0}^{t-1} f_j$, where $\bar{\gamma} = N^{-1}\sum_{i=1}^{N}\gamma_i$. Pesaran (2008) showed that the distribution of the regression t-statistic for $\phi_i = 0$ is free of nuisance parameters. To test the unit root hypothesis in a heterogenous panel the average of the N individual CADF t-statistics (or suitably truncated version of them) can be used. Coakley, Kellard and Smaith (2005) apply the CADF test to real exchange rates of 15 OECD countries.

9.3.4 Other Approaches

A similar approach was proposed by Moon and Perron (2004) and Phillips and Sul (2003a). The test of Moon and Perron (2004) is based on a principal components estimator of $m < N$ common factors $f_{1t}, \ldots, f_{mt}$ in (9.34). The number of common factors can be consistently determined by using the information criteria suggested by Bai and Ng (2002). Let $\widehat{\mathbf{V}}_m = [\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_m]$ denote the matrix of m orthonormal eigenvectors associated with m largest eigenvalues of $\mathbf{\Omega}$. The vector of common factors are estimated as

$$\widehat{\mathbf{f}}_t = [\widehat{f}_{1t}, \ldots, \widehat{f}_{mt}]' = \widehat{\mathbf{V}}_m'\Delta\mathbf{y}_t \ .$$

As shown by Bai and Ng (2002), the principal component estimator $\widehat{\mathbf{f}}_t$ yields a consistent estimator of the factor space as $\min(N,T) \to \infty$. Thus, the elements of the vector

$$\left(\mathbf{I}_N - \widehat{\mathbf{V}}_m\widehat{\mathbf{V}}_m'\right)\Delta\mathbf{y}_t \equiv \mathbf{Q}_{\widehat{V}_m}\Delta\mathbf{y}_t \ , \tag{9.39}$$

are consistent estimates of the idiosyncratic components ξ_{it} as $N \to \infty$. Therefore, by assuming that ξ_{it} is i.i.d., the pooled regression t-statistic

$$t_{MP}^{*} = \frac{\sum_{t=1}^{T}\Delta\widetilde{\mathbf{y}}_t'\mathbf{Q}_{\widehat{V}_m}\widetilde{\mathbf{y}}_{t-1}}{\sqrt{\sum_{t=1}^{T}\widetilde{\mathbf{y}}_{t-1}'\mathbf{Q}_{\widehat{V}_m}\widetilde{\mathbf{y}}_{t-1}}} \ .$$

has a standard normal limiting distribution as $(N, T \to \infty)$ and $\lim \inf_{N,T\to\infty} \log N/\log T \to 0$ (cf. Moon and Perron, 2004).

Hassler et al. (2006) suggest a simple correction for error cross-section dependences that can be used to combine the p-values of individual specific unit root tests, assuming that the correlation among the p-values is constant. The authors found that the suggested combination of p-values yields reliable results even in cases where the correlation is different among the cross-section units.

9.4 Cross-Unit Cointegration

As argued by Banerjee, Marcellino and Osbat (2005) panel unit root tests may be severely biased if the panel units are cross-cointegrated, namely if under the null hypothesis (of unit roots) one or more linear combinations of $\mathbf{y}_t$ are stationary. This needs to be distinguished from the case where the errors are cross correlated without necessarily involving cointegration across the cross section units. Under the former two or more cross section units must share at least one common stochastic trend. Such a situation is likely to occur if the PPP hypothesis is examined (cf. Lyhagen, 2000, Banerjee et al., 2005; Wagner, 2007).

The tests proposed by Moon and Perron (2004) and Pesaran (2008) are based on the model

$$y_{it} = (1-\phi_i)\,\mu_i + \phi_i y_{i,t-1} + \gamma_i f_t + \varepsilon_{it}\ , \tag{9.40}$$

Under the unit root hypothesis, $\phi_i = 1$, this equation yields

$$y_{it} = y_{i0} + \gamma_i s_{ft} + s_{it}\ ,$$

where

$$\begin{aligned} s_{ft} &= f_1 + f_2 + \ldots + f_t\ , \\ s_{it} &= \varepsilon_{i1} + \varepsilon_{i2} + \ldots + \varepsilon_{it}\ . \end{aligned}$$

Clearly, under the null hypothesis all cross section units are related to the common stochastic component, s_{ft}, albeit with varying effects, γ_i. This framework rules out cross-unit cointegration as under the null hypothesis there does not exist a linear combination of $y_{1t}, \ldots, y_{Nt}$ that is stationary. Therefore, tests based on (9.40) are designed to test the joint null hypothesis: "All time series are I(1) and not cointegrated".

To allow for cross-unit cointegration Bai and Ng (2004) proposed analyzing the common factors and idiosyncratic components separately. A simple multi-factor example of Bai and Ng framework is given by

$$\begin{aligned} y_{it} &= \mu_i + \boldsymbol{\gamma}_i' \mathbf{g}_t + e_{it}\ , \\ \Delta \mathbf{g}_t &= \boldsymbol{\Lambda} \boldsymbol{g}_{t-1} + \mathbf{v}_t\ , \\ e_{it} &= \rho_i e_{i,t-1} + \varepsilon_{it}\ , \end{aligned}$$

where $\mathbf{g}_t$ is the $m \times 1$ vector of unobserved components, $\mathbf{v}_t$ and ε_{it} are stationary common and individual specific shocks, respectively. Two different sets of null hypotheses are considered: H_0^a :(testing the $I(0)/I(1)$ properties of the common fcators) $\text{Rank}(\boldsymbol{\Lambda}) = r \leq m$, and H_0^b : (panel unit root tests) $\rho_i = 1$, for all i. A test of H_0^a is based on common factors estimated by principal components and cointegration tests are used to determine the number of common trends, $m - r$. Panel unit root tests are then applied to the idiosyncratic components. The null hypothesis that the time series have a unit root is rejected if either the test of the common factors or the test

for the idiosyncratic component cannot reject the null hypothesis of nonstationary components.[8] As has been pointed out by Westerlund (2007a), replacing the unobserved idiosyncratic components by estimates introduces an asymptotic bias when pooling the t-statistic (or p-values) of the panel units which renders the pooled tests in Bai and Ng (2004) asymptotically invalid. However, in a recent paper Bai and Ng (2007) show that pooled panel unit root tests can still be applied to the estimated idiosyncratic components if the tests are based on the pooled estimator of the largest autoregressive root.

The implementation of the Bai and Ng procedure requires estimates of m and r that might require very large N and T (e.g. Gengenbach, Palm and Urbain, 2006a). Note also that the panel structure of the data does not enhance the power of testing H_0^a, which primarily depends on the time dimension. The cross section dimension only helps in getting more precise estimates of the unobserved common factors. As a result, in panels of typical sample sizes we cannot hope to learn much about the order of integration by applying unit root or cointegration tests to the estimated common factors. Since the low power of testing H_0^a shrinks down the overall power of the combined test procedure it is very hard in practice to establish the stationarity of the variables even if N is extremely large.

To allow for short-run and long-run (cross-unit cointegration) dependencies, Chang and Song (2005) suggest a nonlinear instrument variable test procedure. As the nonlinear instruments suggested by Chang (2002) are invalid in the case of cross-unit cointegration panel specific instruments based on the Hermit function of different order are used as nonlinear instruments. Chang and Song (2005) showed that the t-statistic computed from the nonlinear IV statistic are asymptotically standard normally distributed and, therefore, a panel unit statistics against the heterogeneous alternative H_{1b} can be constructed that has an standard normal limiting distribution.

Choi and Chue (2007) employ a subsampling procedure to obtain tests that are robust against a wide range of cross-section dependence such as weak and strong correlation as well as cross-unit cointegration. To this end the sample is grouped into a number of overlapping blocks of b time periods. Using all $(T-b+1)$ possible overlapping blocks, the critical value of the test is estimated by the respective quantile of the empirical distribution of the $(T-b+1)$ test statistics computed. The advantage of this approach is that the null distribution of the test statistic may depend on unknown nuisance parameters. Whenever the test statistics converge in distribution to some limiting null distribution as $T \to \infty$ and N fixed, the subsample critical values converge in probability to the true critical values. Using Monte Carlo simulations Choi and Chue (2007) demonstrate that the size of the subsample test is indeed very robust against various forms of cross-section dependence.

[8] An alternative factor extraction method is suggested by Kapetanios (2007) who also provides detailed Monte Carlo results on the small sample performance of panel unit root tests based on a number of alternative estimates of the unobserved common factors. He shows that the factor-based panel unit root tests tend to perform rather poorly when the unobserved common factor is serially correlated.

9.5 Finite Sample Properties of Panel Unit Root Tests

It has become standard to distinguish first generation panel unit root tests that are based on the assumption of independent cross section units and second generation tests that allow for some kind of cross-section dependence. Maddala and Wu (1999) compared several first generation tests. For the heterogeneous alternative under consideration they found that in most cases the Fisher test (9.20) performs similar or slightly better than the IPS statistic with respect to size and power. The Levin and Lin statistic (in the version of the 1993 paper) performs substantially worse. Similar results are obtained by Choi (2001). Madsen (2003) derived the local power function against homogeneous alternatives under different detrending procedures. Her Monte Carlo simulations support her theoretical findings that the test based on estimating the mean under the null hypothesis (i.e. the initial observation is subtracted from the time series) outperforms tests based on alternative demeaning procedures. Similar findings are obtained by Bond, Nauges and Windmeijer (2002).

Moon and Perron (2004) compare the finite sample powers of alternative tests against the homogeneous alternative. They found that the point-optimal test of Moon, Perron and Phillips (2007) performs best and show that the power of this test is close to the power envelope. Another important finding from these simulation studies is the observation that the power of the test drops dramatically if a time trend is included. This confirms theoretical results on the local power of panel unit root tests derived by Breitung (2000), Ploberger and Phillips (2002) and Moon et al. (2007).

Hlouskova and Wagner (2006) compare a large number of first generation panel unit root tests applied to processes with MA(1) errors. Not surprisingly, all tests are severely biased as the root of the MA process approaches unity. Overall, the tests of Levin et al. (2002) and Breitung (2000) have the smallest size distortions. These tests also perform best against the homogenous alternative, where the autoregressive coefficient is the same for all panel units. Of course this is not surprising as these tests are optimal under homogeneous alternatives. Furthermore, it turns out that the stationarity tests of Hadri (2000) perform very poorly in small samples. This may be due to the fact that asymptotic values for the mean and variances of the KPSS statistics are used, whereas Levin et al. (2002) and IPS (2003) provide values for small T as well.

The relative performance of several second generation tests have been studied by Gutierrez (2003), and Gengenbach et al. (2006a), where the cross-section dependence is assumed to follow a factor structure. The results very much depend on the underlying model. The simulations carried out by Gengenbach et al. (2006a) show that in general, the mean CADF test has better size properties than the test of Moon and Perron (2004), which tends to be conservative in small samples. However the latter test appears to have more power against stationary idiosyncratic components. Since these tests remove the common factors, they will eventually indicate stationary time series in cases where the series are actually nonstationary due to a common stochastic trend. The results of Gengenbach et al. (2006a) also suggest that

the approach of Bai and Ng (2004) is able to cope with this possibility although the power of the unit test applied to the nonstationary component is not very high.

In general, the application of factor models in the case of weak correlation does not yield valid test procedures. Alternative unit root tests that allow for weak cross section dependence are considered in Breitung and Das (2005). They found that the GLS t-statistic may have a severe size bias if T is only slightly larger than N. In these cases Chang's (2004) bootstrap procedure is able to improve the size properties substantially. The robust OLS t-statistic performs slightly worse but outperforms the nonlinear IV test of Chang (2002). However, Monte Carlo simulations recently carried out by Baltagi, Bresson and Pirotte (2007) show that there can be considerable size distortions even in panel unit root tests that allow for weak dependence. Interestingly enough Pesaran's test, which is not designed for weak cross section dependence, tends to be the most robust to spatial type dependence.

9.6 Panel Cointegration: General Considerations

The estimation of long-run relationships has been the focus of extensive research in time series econometrics. In the case of variables on a single cross section unit the existence and the nature of long-run relations are investigated using cointegration techniques developed by Engle and Granger (1987), Johansen (1991, 1995) and Phillips (1991). In this literature residual-based and system approaches to cointegration are advanced. In this section we review the panel counter part of this literature. But before considering the problem of cointegration in a panel a brief overview of the cointegration literature would be helpful.

Consider the n_i time series variables $\mathbf{z}_{it} = (z_{i1t}, z_{i2t}, \ldots, z_{in_it})'$ observed on the ith cross section unit over the period $t = 1, 2, \ldots, T$, and suppose that for each i

$$z_{ijt} \sim I(1), \qquad j = 1, 2, \ldots, n_i \, .$$

Then $\mathbf{z}_{it}$ is said to form one or more cointegrating relations if there are linear combinations of z_{ijt}'s for $j = 1, 2, \ldots, n_i$ that are $I(0)$ i.e. if there exists an $n_i \times r_i$ matrix $(r_i \geq 1)$ such that

$$\underset{r_i \times n_i}{\beta_i'} \quad \underset{n_i \times 1}{\mathbf{z}_{it}} \quad = \quad \underset{r_i \times 1}{\boldsymbol{\xi}_{it}} \quad \sim I(0) \, .$$

r_i denotes the number of cointegrating (or long-run) relations. The residual-based tests are appropriate when $r_i = 1$, and $\mathbf{z}_{it}$ can be partitioned such that $\mathbf{z}_{it} = (y_{it}, \mathbf{x}_{it}')'$ with no cointegration amongst the $k_i \times 1$ $(k_i = n_i - 1)$ variables, $\mathbf{x}_{it}$. The system cointegration approaches are much more generally applicable and allow for $r_i > 1$ and do not require any particular partitioning of the variables in $\mathbf{z}_{it}$.[9] Another main difference between the two approaches is the way the stationary component of

[9] System approaches to cointegration analysis that allow for weakly exogenous (or long-run forcing) variables has been considered in Pesaran, Shin and Smith (2000).

$\boldsymbol{\xi}_{it}$ is treated in the analysis. Most of the residual-based techniques employ non-parametric (spectral density) procedures to model the residual serial correlation in the error correction terms, $\boldsymbol{\xi}_{it}$, whilst vector autoregressions (VAR) are utilized in the development of system approaches.

In panel data models the analysis of cointegration is further complicated by heterogeneity, unbalanced panels, cross section dependence, cross unit cointegration and the N and T asymptotics. But in cases where n_i and N are small such that $\Sigma_{i=1}^{N} n_i$ is less than 10, and T is relatively large ($T > 100$), as noted by Banerjee, Marcellino and Osbat (2004), many of these problems can be avoided by applying the system cointegration techniques to the pooled vector, $\mathbf{z}_t = (\mathbf{z}'_{1t}, \mathbf{z}'_{2t}, \ldots, \mathbf{z}'_{Nt})'$. In this setting cointegration will be defined by the relationships $\beta' \mathbf{z}_t$ that could contain cointegration between variables from different cross section units as well as cointegration amongst the different variables specific to a particular cross section unit. This framework can also deal with residual cross section dependence since it allows for a general error covariance matrix that covers all the variables in the panel.

Despite its attractive theoretical features, the 'full' system approach to panel cointegration is not feasible even in the case of panels with moderate values of N and n_i. See Sect. 9.10 below for further details. In practice, cross section cointegration can be accommodated using common factors as in the work of Bai and Ng (2004), Pesaran (2006), Pesaran, Schuermann and Weiner (2004, PSW) and its subsequent developments in Dees, di Mauro, Pesaran and Smith (2007, DdPS). Bai and Ng (2004) consider the simple case where $n_i = 1$ but allow N and T to be large. But their set up can be readily generalized so that cointegration within each cross section unit as well as across the units can be considered. Following DdPS suppose that[10]

$$\mathbf{z}_{it} = \boldsymbol{\Gamma}_{id} \mathbf{d}_t + \boldsymbol{\Gamma}_{if} \mathbf{f}_t + \boldsymbol{\xi}_{it} \ , \tag{9.41}$$

for $i = 1, \ldots, N$; $t = 1, 2, \ldots, T$, and to simplify the exposition assume that $n_i = n$, where as before $\mathbf{d}_t$ is the $s \times 1$ vector of deterministics $(1, t)$ or observed common factors such as oil prices, $\mathbf{f}_t$ is a $m \times 1$ vector of unobserved common factors, $\boldsymbol{\Gamma}_{id}$ and $\boldsymbol{\Gamma}_{if}$ are $n \times s$ and $n \times m$ associated unknown coefficient matrices, $\boldsymbol{\xi}_{it}$ is an $n \times 1$ vector of error terms.

Unit root and cointegration properties of $\mathbf{z}_{it}$, $i = 1, 2, \ldots, N$, can be analyzed by allowing the common factors, $\mathbf{f}_t$, and/or the country-specific factors, $\boldsymbol{\xi}_{it}$, to have unit roots. To see this suppose

$$\Delta \mathbf{f}_t = \boldsymbol{\Lambda}(L) \boldsymbol{\eta}_t, \boldsymbol{\eta}_t \sim IID(\mathbf{0}, \mathbf{I}_m) \ , \tag{9.42}$$

$$\Delta \boldsymbol{\xi}_{it} = \boldsymbol{\Psi}_i(L) \mathbf{v}_{it}, \ \ \mathbf{v}_{it} \sim IID(\mathbf{0}, \mathbf{I}_n) \ , \tag{9.43}$$

where L is the lag operator and

[10] DdPS also allow for common observed macro factors (such as oil prices), but they are not included to simplify the exposition.

$$\mathbf{\Lambda}(L) = \sum_{\ell=0}^{\infty} \underset{m\times m}{\mathbf{\Lambda}_\ell} L^\ell, \quad \mathbf{\Psi}_i(L) = \sum_{\ell=0}^{\infty} \underset{n\times n}{\mathbf{\Psi}_{i\ell}} L^\ell . \tag{9.44}$$

The coefficient matrices, $\mathbf{\Lambda}_\ell$ and $\mathbf{\Psi}_{i\ell}$, $i = 1, 2, \ldots, N$, are absolute summable, so that $\text{Var}(\Delta \mathbf{f}_t)$ and $\text{Var}(\Delta \boldsymbol{\xi}_{it})$ are bounded and positive definite, and $[\mathbf{\Psi}_i(L)]^{-1}$ exists. In particular we require that

$$||\sum_{\ell=0}^{\infty} \mathbf{\Psi}_{i\ell}\mathbf{\Psi}'_{i\ell}|| \leq K < \infty , \tag{9.45}$$

where K is a fixed constant.

Using the familiar decomposition

$$\mathbf{\Lambda}(L) = \mathbf{\Lambda}(1) + (1-L)\mathbf{\Lambda}^*(L), \text{ and } \mathbf{\Psi}_i(L) = \mathbf{\Psi}_i(1) + (1-L)\mathbf{\Psi}_i^*(L) ,$$

the common stochastic trend representations of (9.42) and (9.43) can now be written as

$$\mathbf{f}_t = \mathbf{f}_0 + \mathbf{\Lambda}(1)\mathbf{s}_t + \mathbf{\Lambda}^*(L)(\boldsymbol{\eta}_t - \boldsymbol{\eta}_0) ,$$

and

$$\boldsymbol{\xi}_{it} = \boldsymbol{\xi}_{i0} + \mathbf{\Psi}_i(1)\mathbf{s}_{it} + \mathbf{\Psi}_i^*(L)(\mathbf{v}_{it} - \mathbf{v}_{i0}) ,$$

where

$$\mathbf{s}_t = \sum_{j=1}^{t} \boldsymbol{\eta}_j \text{ and } \mathbf{s}_{it} = \sum_{j=1}^{t} \mathbf{v}_{ij} .$$

Using the above results in (9.41) now yields

$$\begin{aligned} \mathbf{z}_{it} = {} & \mathbf{a}_i + \mathbf{\Gamma}_{id}\mathbf{d}_t + \mathbf{\Gamma}_{if}\mathbf{\Lambda}(1)\mathbf{s}_t + \mathbf{\Psi}_i(1)\mathbf{s}_{it} \\ & + \mathbf{\Gamma}_{if}\mathbf{\Lambda}^*(L)\boldsymbol{\eta}_t + \mathbf{\Psi}_i^*(L)\mathbf{v}_{it} , \end{aligned}$$

where[11]

$$\mathbf{a}_i = \mathbf{\Gamma}_{if}[\mathbf{f}_0 - \mathbf{\Lambda}^*(L)\boldsymbol{\eta}_0] + \boldsymbol{\xi}_{i0} - \mathbf{\Psi}_i^*(L)\mathbf{v}_{i0} .$$

In this representation $\mathbf{\Lambda}(1)\mathbf{s}_t$ and $\mathbf{\Psi}_i(1)\mathbf{s}_{it}$ can be viewed as common global and individual-specific stochastic trends, respectively; whilst $\mathbf{\Lambda}^*(L)\boldsymbol{\eta}_t$ and $\mathbf{\Psi}_i^*(L)\mathbf{v}_{it}$ are the common and individual-specific stationary components. From this result it is clear that, in general, it will not be possible to simultaneously eliminate the two types of the common stochastic trends (global and individual-specific) in $\mathbf{z}_{it}$.

Specific cases of interest where it would be possible for $\mathbf{z}_{it}$ to form a cointegrating vector are when $\mathbf{\Lambda}(1) = \mathbf{0}$ or $\mathbf{\Psi}_i(1) = \mathbf{0}$. Under the former panel cointegration exists if $\mathbf{\Psi}_i(1)$ is rank deficient. The number of cointegrating relations could differ across i and is given by $r_i = n - \text{Rank}[\mathbf{\Psi}_i(1)]$. Note that even in this case $\mathbf{z}_{it}$ can be cross-sectionally correlated through the common stationary components, $\mathbf{\Lambda}^*(L)\boldsymbol{\eta}_t$. Under $\mathbf{\Psi}_i(1) = \mathbf{0}$ for all i with $\mathbf{\Lambda}(1) \neq \mathbf{0}$, we will have panel cointegration if there exists

[11] In usual case where $\mathbf{d}_t$ is specified to include an intercept, 1, $\mathbf{a}_i$ can be absorbed into the deterministics.

$n \times r_i$ matrices β_i such that $\beta_i' \mathbf{\Gamma}_{if} \mathbf{\Lambda}(1) = \mathbf{0}$. Notice that this does not require $\mathbf{\Lambda}(1)$ to be rank deficient.

Turning to the case where $\mathbf{\Lambda}(1)$ and $\mathbf{\Psi}_i(1)$ are both non-zero, panel cointegration could still exist but must involve both $\mathbf{z}_{it}$ and $\mathbf{f}_t$. But since $\mathbf{f}_t$ is unobserved it must be replaced by a suitable estimate. The global VAR (GVAR) approach of Pesaran et al. (2004) and Dees et al. (2007) implements this idea by replacing $\mathbf{f}_t$ with the (weighted) cross section averages of $\mathbf{z}_{it}$. To see how this can be justified first differencing (9.41) and using (9.43) note that

$$[\mathbf{\Psi}_i(L)]^{-1}(1-L)\left(\mathbf{z}_{it} - \mathbf{\Gamma}_{id}\mathbf{d}_t - \mathbf{\Gamma}_{if}\mathbf{f}_t\right) = \mathbf{v}_{it} .$$

Using the approximation

$$(1-L)[\mathbf{\Psi}_i(L)]^{-1} \approx \sum_{\ell=0}^{p} \mathbf{\Phi}_{i\ell} L^{\ell} = \mathbf{\Phi}_i(L, p) ,$$

we obtain the following approximate VAR(p) model

$$\mathbf{\Phi}_i(L, p)\left(\mathbf{z}_{it} - \mathbf{\Gamma}_{id}\mathbf{d}_t - \mathbf{\Gamma}_{if}\mathbf{f}_t\right) \approx \mathbf{v}_{it} . \tag{9.46}$$

When the common factors, $\mathbf{f}_t$, are observed the model for the i^{th} cross-section unit decouples from the rest of the units and can be estimated using the econometric techniques developed in Pesaran et al. (2000) with $\mathbf{f}_t$ treated as weakly exogenous. But in general where the common factors are unobserved appropriate proxies for the common factors can be used. There are two possible approaches, one could either use the principal components of the observables, $\mathbf{z}_{it}$, or alternatively following Pesaran (2006) $\mathbf{f}_t$ can be approximated in terms of $\bar{\mathbf{z}}_t = N^{-1}\Sigma_{i=1}^{N}\mathbf{z}_{it}$, the cross section averages of the observables. To see how this procedure could be justified in the present context, average the individual equations given by (9.41) over i to obtain

$$\bar{\mathbf{z}}_t = \bar{\mathbf{\Gamma}}_d \mathbf{d}_t + \bar{\mathbf{\Gamma}}_f \mathbf{f}_t + \bar{\boldsymbol{\xi}}_t , \tag{9.47}$$

where $\bar{\mathbf{\Gamma}}_d = N^{-1}\Sigma_{i=1}^{N}\mathbf{\Gamma}_{id}$, $\bar{\mathbf{\Gamma}}_f = N^{-1}\Sigma_{i=1}^{N}\mathbf{\Gamma}_{if}$, and $\bar{\boldsymbol{\xi}}_t = N^{-1}\Sigma_{i=1}^{N}\boldsymbol{\xi}_{it}$. Also, note from (9.43) that

$$\bar{\boldsymbol{\xi}}_t - \bar{\boldsymbol{\xi}}_{t-1} = N^{-1}\sum_{j=1}^{N}\mathbf{\Psi}_j(L)\mathbf{v}_{jt} . \tag{9.48}$$

But using results in Pesaran (2006), for each t and as $N \to \infty$ we have $\bar{\boldsymbol{\xi}}_t - \bar{\boldsymbol{\xi}}_{t-1} \overset{q.m.}{\to} 0$, and hence $\bar{\boldsymbol{\xi}}_t \overset{q.m.}{\to} \bar{\boldsymbol{\xi}}$, where $\bar{\boldsymbol{\xi}}$ is a time-invariant random variable. Using this result in (9.47) and assuming that the $n \times m$ average factor loading coefficient matrix, $\bar{\mathbf{\Gamma}}_f$, has full column rank (with $n \geq m$) we obtain

$$\mathbf{f}_t \overset{q.m.}{\to} \left(\bar{\mathbf{\Gamma}}_f' \bar{\mathbf{\Gamma}}_f\right)^{-1} \bar{\mathbf{\Gamma}}_f \left(\bar{\mathbf{z}}_t - \bar{\mathbf{\Gamma}}_d \mathbf{d}_t - \bar{\boldsymbol{\xi}}\right) ,$$

which justifies using the observable vector $\{\mathbf{d}_t, \bar{\mathbf{z}}_t\}$ as proxies for the unobserved common factors.

The various contributions to the panel cointegration literature will now be reviewed in the context of the above general set up. First generation literature on panel cointegration tends to ignore the possible effects of global unobserved common factors, or attempts to account for them either by cross-section de-meaning or by using observable common effects such as oil prices or U.S. output. This literature also focusses on residual based approaches where it is often assumed that there exists at most one cointegrating relation in the individual specific models. Notable contributions to this strand of the literature include Kao (1999); Pedroni (1999, 2001, 2004); and more recently Westerlund (2005a). System approaches to panel cointegration that allow for more than one cointegrating relations include the work of Larsson, Lyhagen and Lothgren (2001), Groen and Kleibergen (2003) and Breitung (2005) who generalized the likelihood approach introduced in Pesaran, Shin and Smith (1999). Like the second generation panel unit root tests, recent contributions to the analysis of panel cointegration have also emphasized the importance of allowing for cross section dependence which, as we have noted above, could be due to the presence of common stationary or non-stationary components or both. The importance of allowing for the latter has been emphasized in Banerjee et al. (2004) through the use of Monte Carlo experiments in the case of panels where N is very small, at most 8 in their analysis. But to date a general approach that is capable of addressing all the various issues involved does not exist if N is relatively large.

We now consider in some further detail the main contributions, beginning with a brief discussion of the spurious regression problem in panels.

9.7 Residual-Based Approaches to Panel Cointegration

Under this approach $\mathbf{z}_{it}$ is partitioned as $\mathbf{z}_{it} = (y_{it}, \mathbf{x}'_{it})'$ and the following regressions

$$y_{it} = \delta'_i \mathbf{d}_{it} + \mathbf{x}'_{it}\beta + u_{it}, \; i = 1, 2, \ldots, N, \tag{9.49}$$

are considered, where as before $\delta'_i \mathbf{d}_{it}$ represent the deterministics and the $k \times 1$ vector of regressors, $\mathbf{x}_{it}$, are assumed to be $I(1)$ and not cointegrated. However, the innovations in $\Delta\mathbf{x}_{it}$, denoted by $\varepsilon_{it} = \Delta\mathbf{x}_{it} - E(\Delta\mathbf{x}_{it})$, are allowed to be correlated with u_{it}. Residual-based approaches to panel cointegration focus on testing for unit roots in OLS or panel estimates of u_{it}.

9.7.1 Spurious Regression

Let $\mathbf{w}_{it} = (u_{it}, \varepsilon'_{it})'$ and assume that the conditions for the functional central limit theorem are satisfied such that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{[\cdot T]} \mathbf{w}_{it} \xrightarrow{d} \Sigma_i^{1/2} \mathbf{W}_i(\cdot) \,,$$

where $\mathbf{W}_i$ is a $(k+1) \times 1$ vector of standard Brownian motions, $\xrightarrow{d}$ denotes weak convergence on $D[0,1]$ and

$$\Sigma_i = \begin{bmatrix} \sigma_{i,u}^2 & \boldsymbol{\sigma}_{i,u\varepsilon} \\ \boldsymbol{\sigma}_{i,u\varepsilon}' & \Sigma_{i,\varepsilon\varepsilon} \end{bmatrix} .$$

Kao (1999) showed that in the homogeneous case with $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$, $i = 1, \ldots, N$, and abstracting from the deterministics, the OLS estimator $\widehat{\beta}$ converges in probability to the limit $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1} \boldsymbol{\sigma}_{\varepsilon u}$, where it is assumed that $\mathbf{w}_{it}$ is i.i.d. across i. In the heterogeneous case $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ and $\boldsymbol{\sigma}_{\varepsilon u}$ are replaced by the means $\overline{\Sigma}_{\varepsilon\varepsilon} = N^{-1} \sum_{i=1}^{N} \Sigma_{i,\varepsilon\varepsilon}$ and $\overline{\boldsymbol{\sigma}}_{\varepsilon u} = N^{-1} \sum_{i=1}^{N} \boldsymbol{\sigma}_{i,\varepsilon u}$, respectively (cf. Pedroni, 2000). In contrast, the OLS estimator of β fails to converge within a pure time series framework (cf. Phillips 1987). On the other hand, if $\mathbf{x}_{it}$ and y_{it} are independent random walks, then the t-statistics for the hypothesis that one component of β is zero is $O_p(T^{1/2})$ and, therefore, the t-statistic has similar properties as in the time series case. As demonstrated by Entorf (1997) and Kao (1999), the tendency for spuriously indicating a relationship among y_{it} and $\mathbf{x}_{it}$ may even be stronger in panel data regressions as in the pure time series case. Therefore, it is important to test whether the errors in a panel data regression like (9.49) are stationary.

9.7.2 Tests of Panel Cointegration

As in the pure time series framework, the variables in a regression function can be tested against cointegration by applying unit roots tests of the sort suggested in the previous sections to the residuals of the estimated regression. Unfortunately, panel unit root tests cannot be applied to the residuals in (9.49) if $\mathbf{x}_{it}$ is (long-run) endogenous, that is, if $\boldsymbol{\sigma}_{\varepsilon u} \neq 0$. Letting $T \to \infty$ be followed by $N \to \infty$, Kao (1999) show that the limiting distribution of the DF t-statistic applied to the residuals of a pooled OLS regression of (9.49) is

$$(t_\phi - \sqrt{N}\mu_K)/\sigma_K \xrightarrow{d} \mathcal{N}(0,1) \,, \tag{9.50}$$

where the values of μ_K and σ_K depend on the kind of deterministics included in the regression, the contemporaneous covariance matrix $E(\mathbf{w}_{it}\mathbf{w}_{it}')$ and the long-run covariance matrix Σ_i. Kao (1999) proposed adjusting t_ϕ by using consistent estimates of μ_K and σ_K, where he assumes that the nuisance parameters are the same for all panel units (homogenous short-run dynamics).

Pedroni (2004) suggest two different test statistics for the models with heterogeneous cointegration vectors. Let $\widehat{u}_{it} = y_{it} - \widehat{\boldsymbol{\delta}}_i' \mathbf{d}_{it} - \widehat{\beta}_i' \mathbf{x}_{it}$ denote the OLS residual of

the cointegration regression. Pedroni considers two different classes of test statistics: (i) the "panel statistic" that is equivalent to the unit root statistic against homogeneous alternatives and (ii) the "Group Mean statistic" that is analogous to the panel unit root tests against heterogeneous alternatives. The two versions of the t statistic are defined as

$$\text{panel} \quad Z_{Pt} = \left(\tilde{\sigma}_{NT}^2 \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{u}_{i,t-1}^2 \right)^{-1/2} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{u}_{i,t-1} \hat{u}_{it} - T \sum_{i=1}^{N} \hat{\lambda}_i \right)$$

$$\text{group-mean} \quad \widetilde{Z}_{Pt} = \sum_{i=1}^{N} \left(\hat{\sigma}_{ie}^2 \sum_{t=1}^{T} \hat{u}_{i,t-1}^2 \right)^{-1/2} \left(\sum_{t=1}^{T} \hat{u}_{i,t-1} \hat{u}_{it} - T \hat{\lambda}_i \right)$$

where $\hat{\lambda}_i$ is a consistent estimator of the one-sided long run variance $\lambda_i = \sum_{j=1}^{\infty} E(e_{it} e_{i,t-j})$, $e_{it} = u_{it} - \delta_i u_{i,t-1}$, $\delta_i = E(u_{it} u_{i,t-1}) / E(u_{i,t-1}^2)$, $\hat{\sigma}_{ie}^2$ denotes the estimated variance of e_{it} and $\tilde{\sigma}_{NT}^2 = N^{-1} \sum_{i=1}^{N} \hat{\sigma}_{ie}^2$. Pedroni presents values of μ_p, σ_p^2 and $\tilde{\mu}_p$, $\tilde{\sigma}_p^2$ such that $(Z_{Pt} - \mu_p \sqrt{N}) / \sigma_p$ and $(\widetilde{Z}_{Pt} - \tilde{\mu}_p \sqrt{N}) / \tilde{\sigma}_p$ have standard normal limiting distributions under the null hypothesis.

Other residual-based panel cointegration tests include the recent contribution of Westerlund (2005a) that are based on variance ratio statistics and do not require corrections for the residual serial correlations.

The finite sample properties of some residual based test for panel cointegration are discussed in Baltagi and Kao (2000). Gutierrez (2003) compares the power of various panel cointegration test statistics. He shows that in homogeneous panels with a small number of time periods Kao's tests tend to have higher power than Pedroni's tests, whereas in panels with large T the latter tests performs best. Both test outperform the system test suggested by Larssen et al. (2001). Hlouskova and Wagner (2007) compare various panel cointegration tests in a large scale simulation study. They found that Pedroni's (2004) test based on ADF regressions performs best, whereas all other tests tend to be severely undersized and have very low power in may cases. Furthermore, the system tests suffer from large small sample distortions and are unreliable tools for finding out the correct cointegration rank. Gengenbach et al. (2006b) investigate the performance of Pedroni's tests in cross-dependent models with a factor structure.

9.8 Tests for Multiple Cointegration

It is also possible to adapt Johansen's (1995) multivariate test based on a VAR representation of the variables. Let $\Lambda_i(r)$ denote the cross-section specific likelihood-ratio ("trace") statistic of the hypothesis that there are (at most) r stationary linear combinations in the cointegrated VAR system given by $\mathbf{z}_{it} = (y_{it}, \mathbf{x}_{it}')'$. Following the unit root test proposed in IPS (2003), Larsson et al. (2001) suggested the standardized LR-bar statistic

$$\widetilde{\Lambda}(r) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\Lambda_i(r) - E[\lambda_i(r)]}{\sqrt{\mathrm{Var}[\lambda_i(r)]}} ,$$

to test the null hypothesis that $r = 0$ against the alternative that at most $r = r^0 \geq 1$. Using a sequential limit theory it can be shown that $\widetilde{\Lambda}(r)$ is asymptotically standard normally distributed. Asymptotic values of $E[\lambda_i(r)]$ and $\mathrm{Var}[\lambda_i(r)]$ are tabulated in Larsson et al. (2001) for the model without deterministic terms and Breitung (2005) for models with a constant and a linear time trend. Unlike the residual-based tests, the LR-bar test allows for the possibility of multiple cointegration relations in the panel.

It is also possible to test the null hypothesis that the errors of the cointegration regression are stationary. That is, under the null hypothesis it is assumed that $y_{it}, \mathbf{x}_{it}$ are cointegrated with cointegration rank $r = 1$. McCoskey and Kao (1998) suggest a panel version of Shin's (1994) cointegration test based on the residuals of a fully modified OLS regression. Westerlund (2005b) suggests a related test procedure based on the CUSUM statistic.

9.9 Estimation of Cointegrating Relations in Panels

9.9.1 Single Equation Estimators

First, we consider a single-equation framework where it is assumed that y_{it} and the $k \times 1$ vector of regressors $\mathbf{x}_{it}$ are $I(1)$ with at most one cointegrating relations amongst them, namely that there exists a linear relationship of the form (9.49) such that the error u_{it} is stationary. As before it is assumed that $\mathbf{z}_{it} = (y_{it}, \mathbf{x}'_{it})'$ is i.i.d. across i, and the regressors, $\mathbf{x}_{it}$, are not cointegrated. We do not explicitly consider deterministic terms like individual specific constants or trends as the asymptotic theory applies to mean- or trend-adjusted variables as well.

It is assumed that the vector of coefficients, β, is the same for all cross-section units, that is, a homogeneous cointegration relationship is assumed. Alternatively, it may be assumed that the cointegration parameters are cross section specific (heterogenous cointegration).

By applying a sequential limit theory it can be shown that the OLS estimator of β is $T\sqrt{N}$ consistent and, therefore, the time series dimension is more informative on the long-run coefficients than the cross-section dimension. Furthermore, is important to notice that – as in the time series framework – the OLS estimator is consistent but inefficient in the model with endogenous regressors.

Pedroni (1995) and Phillips and Moon (1999, p. 1085) proposed a "fully-modified OLS" (FM-OLS) approach to obtain an asymptotically efficient estimator for homogenous cointegration vectors. This estimator adjusts for the effects of endogenous regressors and short-run dynamics of the errors (cf. Phillips and Hansen, 1990). To correct for the effect of (long-run) endogeneity of the regressors,

the dependent variable is adjusted for the part of the error that is correlated with the regressor

$$y_{it}^{+} = y_{it} - \boldsymbol{\sigma}_{i,\varepsilon u}' \boldsymbol{\Sigma}_{i,\varepsilon\varepsilon}^{-1} \Delta \mathbf{x}_{it} \ . \tag{9.51}$$

A second correction is necessary when computing the OLS estimator

$$\widehat{\beta}_{FM} = \left[\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x}_{it} \mathbf{x}_{it}' \right]^{-1} \left[\sum_{i=1}^{N} \sum_{t=1}^{T} (\mathbf{x}_{it} y_{it}^{+} - \lambda_{i,\varepsilon u}) \right] \ , \tag{9.52}$$

where

$$\boldsymbol{\lambda}_{i,\varepsilon u} = E \left(\sum_{j=0}^{\infty} \varepsilon_{i,t-j} u_{it} \right) \ .$$

The nuisance parameters can be estimated consistently using familiar nonparametric procedures.

An alternative approach is the "Dynamic OLS" (DOLS) estimator suggested by Saikkonen (1991). This estimator is based on the error decomposition

$$u_{it} = \sum_{k=-\infty}^{\infty} \gamma_k' \Delta \mathbf{x}_{i,t+k} + v_{it} \ , \tag{9.53}$$

where v_{it} is orthogonal to all leads and lags of $\Delta \mathbf{x}_{it}$. Inserting (9.53) in the regression (9.49) yields

$$y_{it} = \beta' \mathbf{x}_{it} + \sum_{k=-\infty}^{\infty} \gamma_k' \Delta \mathbf{x}_{i,t+k} + v_{it} \ . \tag{9.54}$$

In practice the infinite sums are truncated at some small numbers of leads and lags (cf. Kao and Chiang, 2000, Mark and Sul, 2003). Westerlund (2005c) considers data dependent choices of the truncation lags. Kao and Chiang (2000) show that in the homogeneous case with $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ and individual specific intercepts the limiting distribution of the DOLS estimator $\widehat{\beta}_{\text{DOLS}}$ is given by

$$T\sqrt{N}(\widehat{\beta}_{\text{DOLS}} - \beta) \ \xrightarrow{d} \ \mathcal{N}(0, 6\, \sigma_{u|\varepsilon}^2 \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1}) \ ,$$

where

$$\sigma_{u|\varepsilon}^2 = \sigma_u^2 - \sigma_{\varepsilon u}' \boldsymbol{\Sigma}_{\varepsilon\varepsilon}^{-1} \sigma_{\varepsilon u} \ .$$

Furthermore, the FM-OLS estimator possesses the same asymptotic distribution as the DOLS estimator. In the heterogeneous case $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ and $\sigma_{u|\varepsilon}^2$ are replaced by $\overline{\boldsymbol{\Sigma}}_{\varepsilon\varepsilon} = N^{-1} \sum_{i=1}^{N} \boldsymbol{\Sigma}_{i,\varepsilon\varepsilon}$ and $\overline{\sigma}_{u|\varepsilon}^2 = N^{-1} \sum_{i=1}^{N} \sigma_{i,u|\varepsilon}^2$, respectively (cf. Phillips and Moon, 1999). Again, the matrix Σ_i can be estimated consistently (for $T \to \infty$) by using a nonparametric approach.

In many applications the number of time periods is smaller than 20 and, therefore, the kernel based estimators of the nuisance parameters may perform poorly in such

small samples. Pesaran et al. (1999) adapted a parametric model to estimate the cointegration vector based on the error correction format

$$\Delta y_{it} = \phi_i y_{i,t-1} + \gamma_i' \mathbf{x}_{it} + v_{it} \ , \tag{9.55}$$

where for simplicity of exposition we have abstracted from deterministics and lagged changes in y_{it} and $\mathbf{x}_{it}$.[12] It is assumed that the long-run parameters are identical across the cross section units, i.e., $\beta_i = -\gamma_i/\phi_i = \beta$ for $i = 1, \ldots, N$. Economic theory often predicts the same cointegration relation(s) across the cross section units, although is often silent on the magnitude of short-run dynamics, ϕ_i, across i. For example, the long-run relationships predicted by the PPP, the uncovered interest parity, or the Fisher equation are the same across countries, although the speed of convergence to these long-run relations could differ markedly over countries due to differences in economic and political institutions.[13] For further discussions see, for example, Pesaran (1997).

Letting $\zeta_{it}(\beta) = y_{i,t-1} - \beta' \mathbf{x}_{it}$, the model is rewritten as

$$\Delta\, y_{it} = \phi_i \zeta_{it}(\beta) + v_{it} \ . \tag{9.56}$$

Pesaran et al. (1999) have suggested an ML estimation method based on the concentrated likelihood function

$$L_c(\beta) = c - \sum_{i=1}^{N} \frac{T}{2} \log |\widehat{\sigma}^2_{i,v}(\beta)| \ , \tag{9.57}$$

where c is a constant, and

$$\widehat{\sigma}^2_{i,v}(\beta) = \frac{1}{T} \sum_{t=1}^{T} \widetilde{v}_{it}(\beta)^2,$$

$$\widetilde{v}_{it}(\beta) = \Delta y_{it} - \left(\frac{\sum_{t=1}^{T} \Delta y_{it}\, \zeta_{it}(\beta)}{\sum_{t=1}^{T} \zeta_{it}(\beta)^2} \right) \zeta_{it}(\beta) \ .$$

Pesaran et al. (1999) suggested a Gauss–Newton algorithm to maximize (9.57). The means of the error correction coefficients are estimated by the simple average of the individual coefficients (or the ML estimates) of ϕ_i, $i = 1, \ldots, N$. This estimator is called the *pooled mean group estimator*.

[12] Since there are no restrictions on the additional variables they can be concentrated out from the likelihood function by replacing Δy_{it}, $y_{i,t-1}$ and $\mathbf{x}_{it}$ by residuals obtained from regressions on the deterministic terms and lagged differences of the variables, $\Delta\, \mathbf{z}_{i,t-1}, \Delta\, \mathbf{z}_{i,t-2}, \ \ldots$

[13] The problem of testing the slope homogeneity hypothesis in panels is reviewed in Hsiao and Pesaran (2008).

9.9.2 System Estimators

The single equation estimators have several drawbacks that can be avoided by using a system approach. First, these estimators assume that all regressors are I(1) and not cointegrated. If there are more than one cointegration relationships, then the matrix $\mathbf{\Sigma}_{\varepsilon\varepsilon}$ is singular and the asymptotic results are no longer valid. Second, the cointegration relationship has to be normalized such that the variable y_{it} enters with unit coefficient. As has been argued by Boswijk (1995), this normalization is problematical if the original coefficient of the variable y_{it} tends to zero.

In the case of short panels with T fixed and N large, Binder, Hsiao and Pesaran (2005) consider estimation and inference in panel vector autoregressions (PVARs) with homogeneous slopes where (i) the individual effects are either random or fixed, (ii) the time-series properties of the model variables are unknown a priori and may feature unit roots and cointegrating relations. Generalized Method of Moments (GMM) and Quasi Maximum Likelihood (QML) estimators are obtained and compared in terms of their asymptotic and finite sample properties. It is shown that the asymptotic variances of the GMM estimators that are based on levels as well as first-differences of the model variables depend on the variance of the individual effects; whereas by construction the fixed effects QML estimator is not subject to this problem. Monte Carlo evidence is provided showing that the fixed effects QML estimator tends to outperform the various GMM estimators in finite sample under both normal and non-normal errors. The paper also shows how the fixed effects QML estimator can be successfully used for unit root and cointegration tests in short panels.

In the case of panels with large N and T, Larsson and Lyhagen (1999), Groen and Kleibergen (2003) and Breitung (2005) consider the vector error correction model (VECM) for the $k+1$ dimensional vector $\mathbf{z}_{it} = (y_{it}, \mathbf{x}_{it}')'$ given by

$$\Delta \mathbf{z}_{it} = \alpha_i \beta_i' \mathbf{z}_{i,t-1} + \mathbf{w}_{it} \,, \tag{9.58}$$

where $\mathbf{w}_{it} = (u_{it}, \varepsilon_{it}')'$ and once again we leave out deterministic terms and lagged differences. To be consistent with the approaches considered above, we confine ourselves to the case of homogenous cointegration, that is, we let $\beta_i = \beta$ for $i = 1, \ldots, N$. Larsson and Lyhagen (1999) propose a ML estimator, whereas the estimator of Groen and Kleibergen (2003) is based on a nonlinear GMM approach.

It is well known that the ML estimator of the cointegration parameters for a single series may behave poorly in small samples. Phillips (1994) has shown that the finite sample moments of the estimator do not exist. Using Monte Carlo simulations Hansen, Kim and Mittnik (1998) and Brüggemann and Lütkepohl (2005) found that the ML estimator may produce implausible estimates far away from the true parameter values. Furthermore the asymptotic χ^2 distribution of the likelihood ratio test for restrictions on the cointegration parameters may be a poor guide for small sample inference (e.g. Gredenhoff and Jacobson, 2001).

To overcome these problems, Breitung (2005) proposed a computationally convenient two-step estimator, which is adopted from Ahn and Reinsel (1990). This

estimator is based on the fact that the Fisher information is block-diagonal with respect to the short and long-run parameters. Accordingly, an asymptotically efficient estimator can be constructed by estimating the short and long-run parameters in separate steps. Suppose that the $n \times r$ matrix of cointegrating vectors is "normalized" as $\beta = (\mathbf{I}_r, \mathbf{B})'$, where $\mathbf{I}_r$ is the identity matrix of order r and $\mathbf{B}$ is the $(n-r) \times r$ matrix of unknown coefficients.[14] Then β is exactly identified and the Gaussian ML estimator of $\mathbf{B}$ is equivalent to the OLS estimator of $\mathbf{B}$ in

$$\mathbf{z}_{it}^{*} = \mathbf{B}\mathbf{z}_{i,t-1}^{(2)} + \mathbf{v}_{it} \ , \tag{9.59}$$

where $\mathbf{z}_{it}^{(2)}$ is the $r \times 1$ vector defined by $\mathbf{z}_{it} = [\mathbf{z}_{it}^{(1)'}, \mathbf{z}_{it}^{(2)'}]'$, and

$$\mathbf{z}_{it}^{*} = (\boldsymbol{\alpha}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\alpha}_i)^{-1} \boldsymbol{\alpha}_i' \boldsymbol{\Sigma}_i^{-1} \Delta \mathbf{z}_{it} - \mathbf{z}_{i,t-1}^{(1)} \ .$$

The matrices $\boldsymbol{\alpha}_i$ and $\boldsymbol{\Sigma}_i$ can be replaced by $\sqrt{T}$-consistent estimates without affecting the limiting distribution. Accordingly, these matrices can be estimated for each panel unit separately, e.g., by using Johansen's (1991) ML estimator. To obtain the same normalization as in (9.59) the estimator for $\boldsymbol{\alpha}_i$ is multiplied with the $r \times r$ upper block of the ML estimator of β.

Breitung (2005) showed that the limiting distribution of the OLS estimator of $\mathbf{B}$ is asymptotically normal. Therefore, tests of restrictions on the cointegration parameters have the standard limiting distributions (i.e. a χ^2 distribution for the usual Wald tests).

Some Monte Carlo experiments were performed by Breitung (2005) to compare the small sample properties of the two-step estimator with the FM-OLS and DOLS estimators. The results suggest that the latter two tests may be severely biased in small samples, whereas the bias of the two-step estimator is relatively small. Furthermore, the standard errors (and hence the size properties of the t-statistics) of the two-step procedure are more reliable than the ones of the semi-parametric estimation procedures. In a large scale simulation study, Hlouskova and Wagner (2007) found that the DOLS estimator outperforms all other estimators, whereas the FM-OLS and the two-step estimator perform similarly.

9.10 Cross-Section Dependence and the Global VAR

As pointed out earlier an important limitation of the econometric approaches discussed so far is that they assume that all cross-section units are independent. In many applications based on multi-country data sets this assumption is clearly unrealistic.

[14] The analysis can be readily modified to take account of other types of exact identifying restrictions on β that might be more appropriate from the view-point of long-run economic theory. See Pesaran and Shin (2002) for a general discussion of identification and testing of cointegrating relations in the context of a single cross section units.

To accommodate cross-dependence among panel units Mark, Ogaki and Sul (2004) and Moon and Perron (2007) proposed a Dynamic Seemingly Unrelated Regression (DSUR) estimator. Their approach is based on a GLS estimator of the dynamic representation (9.54) when there exists a single cointegrating relation between y_{it} and $\mathbf{x}_{it}$, and does not allow for the possibility of cross unit cointegration. Let $\mathbf{h}_{it}(p) = \left(\Delta\mathbf{x}'_{i,t-p}, \ldots, \Delta\mathbf{x}'_{i,t+p}\right)'$ and $\mathbf{h}_{pt} = (\mathbf{h}_{1t}(p)', \ldots, \mathbf{h}_{Nt}(p)')'$. To correct for endogeneity of the regressors, first y_{it} and $\mathbf{x}_{it}$ are regressed on $\mathbf{h}_{pt}$. Let $\tilde{y}_{it}$ and $\tilde{\mathbf{x}}_{it}$ denote the resulting regression residuals. Furthermore, define $\tilde{\mathbf{y}}_t = (\tilde{y}_{1t}, \ldots, \tilde{y}_{Nt})'$ and $\widetilde{\mathbf{X}}_t = (\tilde{\mathbf{x}}_{1t}, \ldots, \tilde{\mathbf{x}}_{Nt})'$. The DSUR estimator of the (homogeneous) cointegration vector is

$$\widehat{\beta}_{\text{dsur}} = \left(\sum_{t=p+1}^{T-p} \widetilde{\mathbf{X}}_t' \Sigma_{uu}^{-1} \widetilde{\mathbf{X}}_t\right) \sum_{t=p+1}^{T-p} \widetilde{\mathbf{X}}_t' \Sigma_{uu}^{-1} \tilde{\mathbf{y}}_t \tag{9.60}$$

where $\mathbf{\Sigma}_{uu}$ denotes the long-run covariance matrix of $\mathbf{u}_t = (u_{1t}, \ldots, u_{Nt})'$, namely

$$\mathbf{\Sigma}_{uu} = \lim_{T\to\infty} \frac{1}{T} E\left[\left(\sum_{t=1}^{T} \mathbf{u}_t\right)\left(\sum_{t=1}^{T} \mathbf{u}'_t\right)\right],$$

for a fixed N. This matrix is estimated by using an autoregressive representation of $\mathbf{u}_t$. See also (9.53). An alternative approach is suggested by Breitung (2005), where a SUR procedure is applied in the second step of the two-step estimator.

To estimate panel data regression models with a multifactor error structure Pesaran (2006) proposed the common correlated effects (CCE) estimator. The basic idea of this estimation procedure is to filter the individual-specific regressors by means of cross-section averages such that the differential effects of unobserved common factors are eliminated. An extension of this analysis to non-stationary common factors is provided in Kapetanios, Pesaran and Yamagata (2006). Also as shown by Kapetanios (2007) the CCE estimator can be applied to a cointegrated panel data regression. An empirical application to the analysis of house prices in the US is provided in Holly, Pesaran and Yamagata (2007).

Bai and Kao (2005), Westerlund (2007b), and Bai, Kao and Ng (2007) suggest estimators for the cointegrated panel data model given by

$$y_{it} = \beta' \mathbf{x}_{it} + \gamma_i' \mathbf{f}_t + e_{it} \tag{9.61}$$

where f_t is a $r \times 1$ vector of common factors and e_{it} is the idiosyncratic error. Bai and Kao (2005) and Westerlund (2007b) assume that f_t is stationary. They suggest an FM-OLS cointegration regression that accounts for the cross-correlation due to the common factors. Bai et al. (2007) consider a model with nonstationary factors. Their estimation procedure is based on a sequential minimization of the criterion function

$$S_{NT}(\beta, \mathbf{f}_1, \ldots, \mathbf{f}_T, \gamma_1, \ldots, \gamma_N) = \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \beta' \mathbf{x}_{it} - \gamma_i' \mathbf{f}_t)^2 \tag{9.62}$$

subject to the constraint $T^{-2}\sum_{t=1}^{T} f_t f_t' = I_r$ and $\sum_{i=1}^{N} \gamma_i \gamma_i'$ being diagonal. The asymptotic bias of the resulting estimator is corrected for by using an additive bias adjustment term or by using a procedure similar to the FM-OLS estimator suggested by Phillips and Hansen (1990).

A common feature of these approaches is that cross-section dependence can be represented by a contemporaneous correlation of the errors, and do not allow for the possibility of cross unit cointegration. In many applications it is more realistic to allow for some form of dynamic cross-section dependence. A general model to accommodate cross-section cointegration and dynamic links between panel units is the panel VECM model considered by Groen and Kleibergen (2003) and Larsson and Lyhagen (1999). As in Sect. 9.6, let $\mathbf{z}_{it}$ denote a n-dimensional vector of times series on the i^{th} cross section unit. Consider the $nN \times 1$ vector $\mathbf{z}_t = (\mathbf{z}_{1t}', \ldots, \mathbf{z}_{Nt}')'$ of all available time series in the panel data set. The VECM representation of this time series vector is

$$\Delta \mathbf{z}_t = \Pi \mathbf{z}_{t-1} + \Gamma_1 \Delta \mathbf{z}_{t-1} + \cdots + \Gamma_p \Delta \mathbf{z}_{t-p} + \mathbf{u}_t \ . \tag{9.63}$$

For cointegrated systems $\text{rank}(\Pi) < nN$. It is obvious that such systems typically involve a large number of parameters as the number of parameters increases with N^2. Therefore, to obtain reliable estimates of the parameters T must be considerably larger than N. In many macroeconomic applications, however, the number of time periods is roughly as large as the number of cross-section units. Therefore, a simple structure must be imposed on the matrices $\mathbf{\Pi}, \mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_p$ that yields a reasonable approximation to the underlying dynamic system.

The Global VAR (GVAR) introduced by Pesaran, Schuermann and Weiner (2004) and further developed in Dees et al. (2007) can be seen as a theory guided reduction of the general dynamic model given by (9.41), (9.42) and (9.43). In the context of this set up the individual cross section models in the GVAR can be approximated by the VARZ$^*(p_i, p_i)$ in $\mathbf{z}_{it}$ and $\mathbf{z}_{it}^*$:[15]

$$\mathbf{\Phi}_i(L, p_i)\left(\mathbf{z}_{it} - \mathbf{\Gamma}_{id}\mathbf{d}_t - \mathbf{\Gamma}_{if}\mathbf{z}_{it}^*\right) = \mathbf{v}_{it} \ . \tag{9.64}$$

for $i = 1, 2, \ldots, N$, where $\mathbf{d}_t$ are the observed common effects (such as intercepts, time trends or oil prices), and $\mathbf{z}_{it}^*$ is defined by

$$\mathbf{z}_{it}^* = \sum_{j=1}^{N} w_{ij}\mathbf{z}_{jt} \ .$$

The weights, w_{ij}, $j = 1, 2, \ldots, N$ must satisfy the following conditions

$$w_{ii} = 0, \ \sum_{j=1}^{N} w_{ij} = 1, \text{ and } \sum_{j=1}^{N} w_{ij}^2 \to 0, \text{ as } N \to \infty \ ,$$

[15] VARZ* represents a VAR model augmented with $\mathbf{z}_{it}^*$ as weakly exogenous variables.

and could be time varying. Typical examples of such weights are $w_{ij} = 1/(N-1)$, for $i \neq j$, trade weights or other measures of economic distance between the cross section units. The estimation of (9.64) can proceed by treating the cross section averages as weakly exogenous $I(1)$ variables using standard time series cointegration techniques developed, for example, in Pesaran et al. (2000). The assumption that $\mathbf{z}_{it}^{*}$ are weakly exogenous $I(1)$, or long-run forcing, for $\mathbf{z}_{it}$, can be tested. For further details see Pesaran et al. (2004). It turns out that this is a reasonable assumption for all countries except for the U.S. where most of the variables should be treated as endogenous.

9.11 Concluding Remarks

As this review shows the literature on panel unit roots and cointegration has been expanding very rapidly; in part responding to the complex nature of the interactions and dependencies that generally exist over time and across the individual units in the panel. Observations on firms, industries, regions and countries tend to be cross correlated as well as serially dependent. The problem of cross section dependence is particularly difficult to deal with since it could arise for a variety of reasons; spatial spill over effects, common unobserved shocks, social interactions or a combination of these factors. Parameter heterogeneity and deterministics also pose additional difficulties and how they are treated under the null and the alternative hypothesis can affect the outcome of the empirical analysis.

Initially, the panel unit root and cointegration tests were developed assuming that the errors of the individual equations are cross sectionally independent. These, referred to as the first generation tests, continue to form an important part of the literature, providing a theoretical basis for the more recent (second generation) developments that attempt to take account of the residual cross section dependence in panels in the case of panels where the time dimension and the cross section dimension are both relatively large. In the analysis of cointegration the hypothesis testing and estimation problems are also further complicated by the possibility of cross section cointegration. These and other issues are currently the subject of extensive research.

Acknowledgments We are grateful to Jushan Bai, Badi Baltagi, Jaroslava Hlouskova, George Kapetanios, Uwe Hassler, Serena Ng, Elisa Tosetti, Ron Smith, and Joakim Westerlund and an anonymous referee for comments on a preliminary version of this chapter.

References

Ahn, S.K. and G.C. Reinsel (1990), Estimation for Partially Nonstationary Multivariate Autoregressive Models, Journal of the American Statistical Association, 85, 813–823.

Andrews, D.W.K. (1998), Hypothesis Testing with a Restricted Parameter Space, Journal of Econometrics, 84, 155–199.

Arellano, M. (2003), Panel Data Econometrics, Oxford: Oxford University Press.
Bai J. and C. Kao (2005), On the Estimation and Inference of a Panel Cointegration Model with Cross-Sectional Dependence, in: B. Baltagi (Ed.), Contributions to Economic Analysis, Elsevier, Amsterdam).
Bai, J. and S. Ng (2002), Determining the Number of Factors in Approximate Factor Models, Econometrica, 70, 191–221.
Bai, J. and S. Ng (2004), A Panic Attack on Unit Roots and Cointegration, Econometrica, 72, 1127–1177.
Bai, J. and S. Ng (2007), Panel Unit Root Tests with Cross-Section Dependence: A Further Investigation, New York University, unpublished.
Bai, J., C. Kao and S. Ng (2007), Panel Cointegration with Global Stochastic Trends, New York University, unpublished.
Baltagi, B.H. and C. Kao (2000), Nonstationary Panels, Cointegration in Panels and Dynamic Panels, A Survey, in: B. Baltagi (ed.), Nonstationary Panels, Panel Cointegration, and Dynamic Panels, Advances in Econometrics, Vol. 15, JAI Press, Amsterdam, 7–52.
Baltagi, B.H., G. Bresson, and A. Pirotte, (2007), Panel Unit Root Tests and Spatial Dependence, Journal of Applied Econometrics, 22, 339–360.
Banerjee, A. (1999), Panel Data Unit Roots and Cointegration: An Overview, Oxford Bulletin of Economics and Statistics, 61, 607–629.
Banerjee, A., M. Marcellino and C. Osbat (2004), Some Cautions on the Use of Panel Methods for Integrated Series of Macroeconomic Data, Econometrics Journal, 7, 322–340.
Banerjee, A., M. Marcellino and C. Osbat (2005), Testing for PPP: Should we use Panel Methods? Empirical Economics, 30, 77–91.
Binder, M., Hsiao, C. and M. H. Pesaran (2005), Estimation and Inference in Short Panel Vector Autoregressions with Unit Roots and Cointegration, Econometric Theory 21, 795–837.
Bond, S., C. Nauges and F. Windmeijer (2002), Unit Roots and Identification in Autoregressive Panel Data Models: A Comparison of Alternative Tests, unpublished.
Boswijk, H. P. (1995). Efficient Inference on Cointegration Parameters in Structural Error Correction Models, Journal of Econometrics, 69, 133–158.
Breitung, J. (2000), The Local Power of Some Unit Root Tests for Panel Data, in B. Baltagi (ed.), Nonstationary Panels, Panel Cointegration, and Dynamic Panels, Advances in Econometrics, Vol. 15, JAI Press, Amsterdam, 161–178.
Breitung, J. (2002), Nonparametric Tests for Unit Roots and Cointegration, Journal of Econometrics, 108, 343–363.
Breitung, J. (2005), A Parametric Approach to the Estimation of Cointegration Vectors in Panel Data, Econometric Reviews, 151–174.
Breitung J. and B. Candelon (2005), Purchasing Power Parity During Currency Crises: A panel unit root test under structural breaks, World Economic Review, 141, 124–140.
Breitung, J. and S. Das (2005), Panel Unit Root Tests Under Cross Sectional Dependence, Statistica Neerlandica, 59, 414–433.
Breitung, J. and S. Das (2008), Testing for Unit Roots in Panels With a Factor Structure, forthcoming in: Econometric Theory.
Breitung, J. and W. Meyer (1994), Testing for Unit Roots in Panel Data: Are Wages on Different Bargaining Levels Cointegrated? Applied Economics, 26, 353–361.
Brüggemann, R. and H. Lütkepohl (2005), Practical Problems with Reduced Rank ML Estimators for Cointegration Parameters and a Simple Alternative, Oxford Bulletin of Economics and Statistics, 67, 673–690.
Carrion-i-Sevestre, J.L., T. Del Barrio and E. Lopez-Bazo (2005), Breaking the Panels: An Application to the GDP Per Capita, Econometrics Journal, 8, 159–175.
Chang, Y. (2002), Nonlinear IV Unit Root Tests in Panels with Cross-Sectional Dependency, Journal of Econometrics, 110, 261–292.
Chang, Y. (2004), Bootstrap Unit Root Tests in Panels with Cross-Sectional Dependency, Journal of Econometrics, 120, 263–293.

Chang, Y., and J.Y. Park (2004), Taking a New Contour: A New Approach to Panel Unit Root Tests, Rice University, unpublished.
Chang, Y., J.Y. Park and P.C.B. Phillips (2001), Nonlinear Econometric Models with Cointegrated and Deterministically Trending Regressors, Econometrics Journal, 4, 1–36.
Chang, Y. and W. Song (2005), Unit Root Tests for Panels in the Presence of Short-Run and Long-Run Dependencies: Nonlinear IV Approach with Fixed N and Large T, Rice University, unpublished.
Choi, I. (2001), Unit Root Tests for Panel Data, Journal of International Money and Finance, 20, 249–272.
Choi, I. (2002), Combination Unit Root Tests for Cross-Sectionally Correlated Panels, in Econometric Theory and Practice: Frontiers of Analysis and Applied Research, Essays in Honor of P.C.B. Phillips, Cambridge University Press, Cambridge.
Choi, I. (2006), Nonstationary Panels, in K. Patterson and T.C. Mills (eds.) Palgrave Handbooks of Econometrics, Vol. 1, 511–539. Palgrave Macmillan, New York.
Choi, I. and T.K. Chue (2007), Subsampling Hypothesis Tests for Nonstationary Panels with Applications to Exchange Rates and Stock Prices, Journal of Applied Econometrics, 22, 233–264.
Coakley, J. and A.M. Fuertes (1997), New Panel Unit Root Tests of PPP, Economics Letters, 57, 17–22.
Coakley, J., N. Kellard and S. Smaith (2005), The PPP debate: Price Matters!, Economic Letters, 88, 209–213.
Dees, S., F. di Mauro, M. H. Pesaran and L. V. Smith (2007), Exploring the International Linkages of the Euro Area: A Global VAR Analysis, Journal of Applied Econometrics, 22, 1, 1–38.
Dickey, D.A. and W.A. Fuller (1979), Distribution of the Estimates for Autoregressive Time Series With a Unit Root, Journal of the American Statistical Association 74, 427–431.
Elliott, G., T. Rothenberg, and J. Stock (1996), Efficient Tests for an Autoregressive Unit Root, Econometrica, 64, 813–836.
Engle, R.F., and C.W.J. Granger (1987), Co-integration and Error Correction: Representation, Estimation, and Testing, Econometrica, 55, 251–276.
Entorf, H. (1997), Random Walks With Drifts: Nonsense Regression and Spurious Fixed-effects Estimation, Journal of Econometrics, 80, 287–296.
Fisher, R.A. (1932), Statistical Methods for Research Workers, Oliver and Bond, Edinburgh, 4th ed.
Flores, R., P. Jorion, P.Y. Preumont and A. Szarfarz (1999), Multivariate Unit Root Tests of the PPP Hypothesis, Journal of Empirical Finance, 6, 335–353.
Fuller, W.A., (1996), Introduction to Statistical Time Series, Second Edition, Wiley.
Gengenbach, C. F.C. Palm and J.-P. Urbain (2006a), Panel Unit Root Tests in the Presence of Cross-Sectional Dependencies: Comparison and Implications for Modelling, Universiteit Maastricht, unpublished.
Gengenbach, C. F.C. Palm and J.-P. Urbain (2006b), Cointegration Testing in Panels with Common Factors", Oxford Bulletin of Economics and Statistics, 68, 683–719.
Gredenhoff, M. and T. Jacobson (2001), Bootstrap Testing Linear Restrictions on Cointegrating Vectors, Journal of Business and Economic Statistics, 19, 63–72.
Groen, J.J.J. and F. Kleibergen (2003), Likelihood-Based Cointegration Analysis in Panels of Vector Error-Correction Models, Journal of Business and Economic Statistics 21: 295–318.
Gutierrez, L. (2003), On the Power of Panel Cointegration Tests: A Monte Carlo comparison, *Economics Letters*, 80, 105–111.
Gutierrez, L. (2006), Panel Unit Roots Tests for Cross-Sectionally Correlated Panels: A Monte Carlo Comparison, Oxford Bulletin of Economics and Statistics, 68, 519–540.
Hadri, K. (2000), Testing for Stationarity in Heterogeneous Panel Data, Econometrics Journal, 3, 148–161.
Hadri, K. and R. Larsson (2005), Testing for Stationarity in Heterogeneous Panel Data Where the Time Dimension is Fixed, Econometrics Journal, 8, 55–69.
Hansen, G., J.R. Kim and S. Mittnik (1998), Testing Cointegrating Coefficients in Vector Autoregressive Error Correction Models, Economics Letters, 58, 1–5.

Harris, D., S. Leybourne and B. McCabe (2004), Panel Stationarity Tests for Purchasing Power Parity with Cross-sectional Dependence, Journal of Business and Economic Statistics, 23, 395–409.

Harris, R.D.F. and E. Tzavalis (1999), Inference for Unit Roots in Dynamic Panels where the Time Dimension is Fixed, Journal of Econometrics, 91, 201–226.

Harvey, A. and D. Bates (2003), Multivariate Unit Root Tests, Stability and Convergence, University of Cambridge, DAE Working Paper No. 301, University of Cambridge, England.

Harvey, D.I., S.J. Leybourne, and N.D. Sakkas (2006), Panel Unit Root Tests and the Impact of Initial Observations, Granger Centre Discussion Paper No. 06/02, University of Nottingham.

Hassler, U., M. Demetrescu and A. Tarcolea (2006), Combining Significance of Correlated Statistics with Application to Panel Data, Oxford Bulletin of Economics and Statistics, 68, 647–663.

Hlouskova, J. and M. Wagner (2006), The Performance of Panel Unit Root and Stationarity Tests: Results from a Large Scale Simulation Study, Econometric Reviews, 25, pp. 85–116.

Hlouskova, J. and M. Wagner (2007), The Performance of Panel Cointegration Methods: Results from a Large Scale Simulation Study, Institute for Advanced Studies, Vienna, unpublished.

Holly, S., M.H. Pesaran, and T. Yamagata (2007), A Spatial-Temporal Model of House Prices in the US, Cambridge University, unpublished.

Hsiao, C., and M. H. Pesaran (2007), Random Coefficient Panel Data Models, in L. Matyas and P. Sevestre, (eds.), The Econometrics of Panel Data, Kluwer Academic Publishers.

Hurlin, C. and V. Mignon, (2004), Second Generation Panel Unit Root Tests, THEMA-CNRS, University of Paris X, unpublished.

Im, K.S., J. Lee and M. Tieslau (2005), Panel LM Unit Root Tests with Level Shifts, Oxford Bulletin of Economics and Statistics, 63, 393–419.

Im, K.S., M.H. Pesaran, and Y. Shin (1995), Testing for Unit Roots in Heterogenous Panels, DAE Working Papers Amalgamated Series No. 9526, University of Cambridge.

Im, K.S., M.H. Pesaran, and Y. Shin (2003), Testing for Unit Roots in Heterogenous Panels, Journal of Econometrics, 115, 53–74.

Johansen, S. (1991), Estimation and Hypothesis Testing of Cointegrating Vectors in Gaussian Vector Autoregressive Models, Econometrica, 59, 1551–1580.

Johansen, S. (1995), Likelihood-based Inference in Cointegrated Vector Autoregressive Models, Oxford: Oxford University Press.

Jönsson, K. (2005), Cross-Sectional Dependency and Size Distortion in a Small-sample Homogeneous Panel-data Unit Root Test, Oxford Bulletin of Economics and Statistics, 63, 369–392.

Kao, C. (1999), Spurious Regression and Residual-based Tests for Cointegration in Panel Data, Journal of Econometrics, 90, 1–44.

Kao, C. and M.-H. Chiang (2000), On the Estimation and Inference of a Cointegrated Regression in Panel Data, in: Baltagi B. (ed.), Nonstationary Panels, Panel Cointegration, and Dynamic Panels, Advances in Econometrics, Vol. 15, Amsterdam: JAI Press, 161–178.

Kapetanios, G. (2007), Dynamic Factor Extraction of Cross-Sectional Dependence in Panel Unit Root Tests, Journal of Applied Econometrics, 22, 313–338.

Kapetanios, G. and M.H. Pesaran (2007), Alternative Approaches To Estimation And Inference In Large Multifactor Panels: Small Sample Results With An Application To Modelling Of Asset Return", in G. Phillips and E. Tzavalis (eds.), The Refinement of Econometric Estimation and Test Procedures: Finite Sample and Asymptotic Analysis, Cambridge University Press, Cambridge.

Kapetanios, G., M. H. Pesaran and T. Yamagata, (2006), Panels with Nonstationary Multifactor Error Structures, CESifo Working Paper No. 1788.

Kiefer, N. and T. Vogelsang (2002), Heteroskedasticity-Autocorrelation Robust Standard Errors Using the Bartlett Kernel Without Truncation, Econometrica, 70, 2093–2095.

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. and Shin, Y. (1992) Testing the Null Hypothesis of Stationary Against the Alternative of a Unit Root: How Sure are we that Economic Time Series Have a Unit Root? Journal of Econometrics 54, 159–178.

Larsson, R. and J. Lyhagen (1999). Likelihood-based Inference in Multivariate Panel Cointegration Models. Working paper series in Economics and Finance, no. 331, Stockholm School of Economics.

Larsson, R., J. Lyhagen and M. Lothgren (2001), Likelihood-based Cointegration Tests in Heterogenous Panels, Econometrics Journal 4, 109–142.

Levin, A. and C.F. Lin (1993), Unit root tests in panel data: asymptotic and finite-sample properties. Unpublished manuscript, University of California, San Diego.

Levin, A., C. Lin, and C.J. Chu (2002), Unit Root Tests in Panel Data: Asymptotic and Finite-sample Properties, Journal of Econometrics, 108, 1–24.

Leybourne, S.J. (1995), "Testing for Unit Roots using Forward and Reverse Dickey-Fuller Regressions", Oxford Bulletin of Economics and Statistics, 57, 559–571.

Lyhagen, J. (2000), Why not Use Standard Panel Unit Root Tests for Testing PPP, Stockholm School of Economics, unpublished.

Maddala, G.S. and Wu, S. (1999), A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test, Oxford Bulletin of Economics and Statistics, 61, 631–652.

Madsen, E. (2003), Unit Root Inference in Panel Data Models Where the Time-series Dimension is fixed: A comparison of different test, CAM working paper No. 2003–13.

Mark, N.C. and D. Sul (2003), Cointegration Vector Estimation by Panel DOLS and Long-run Money Demand, Oxford Bulletin of Economics and Statistics, 65, 655–680.

Mark, N.C., M. Ogaki and D. Sul (2004), Dynamic Seemingly Unrelated Cointegration Regression, forthcoming in: Review of Economic Studies.

McCoskey, S. and C. Kao (1998), A Residual-Based Test of the Null of Cointegration in Panel Data, Econometric Reviews, 17, 57–84.

Moon, R. and B. Perron (2004), Testing for Unit Root in Panels with Dynamic Factors, Journal of Econometrics, 122, 81–126.

Moon, H.R. and B. Perron (2007), Efficient Estimation of the SUR Cointegration Regression Model and Testing for Purchasing Power Parity, forthcoming in: Econometric Reviews.

Moon, H.R., B. Perron, and P.C.B. Phillips (2006), On the Breitung Test for Panel Unit Roots and Local Asymptotic Power, Econometric Theory, 22, 1179–1190.

Moon, H.R., B. Perron, and P.C.B. Phillips (2007), Incidental Trends and the Power of Panel Unit Root Tests, Journal of Econometrics, 141, 416–459.

Murray, C. J. and D.H. Papell (2002), Testing for Unit Roots in Panels in the Presence of Structural Change With an Application to OECD Unemployment, in B. Baltagi (ed.), Nonstationary Panels, Panel Cointegration, and Dynamic Panels, Advances in Econometrics, Vol. 15, JAI, Amsterdam, 223–238.

Nabeya, S. (1999), Asymptotic Moments of Some Unit Root Test Statistics in the Null Case, Econometric Theory, 15, 139–149.

O'Connell, P. (1998), The Overvaluation of Purchasing Power Parity, Journal of International Economics, 44, 1–19.

Park, H. and W. Fuller (1995), Alternative Estimators and Unit Root Tests for the Autoregressive Process, Journal of Time Series Analysis, 16, 415–429.

Pedroni, P. (1995), Panel Cointegration: Asymptotic and Finite Sample Properties of Pooled Time Series Test with an Application to the PPP Hypothesis, Indiana University Working Papers in Economics, No. 95–013.

Pedroni, P. (1999), Critical Values for Cointegration Tests in Heterogeneous Panels with Multiple Regressors, Oxford Bulletin of Economics and Statistics, 61, 653–670.

Pedroni, P. (2000), Fully Modified OLS for Heterogenous Cointegrated Panels, in: Baltagi B. (ed.), Nonstationary Panels, Panel Cointegration, and Dynamic Panels, Advances in Econometrics, Vol. 15, Amsterdam: JAI Press, pp. 93–130.

Pedroni, P. (2001), Purchasing Power Parity Tests in Cointegrated Panels, Review of Economics and Statistics, 83, 727–731.

Pedroni, P. (2004), Panel Cointegration: Asymptotic and Finite Sample Properties of Pooled Time Series Tests With an Application to the PPP Hypothesis, Econometric Theory, 20, 597–625.

Pedroni, P. and T. Vogelsang (2005), Robust Tests for Unit Roots in Heterogeneous Panels, Williams College, unpublished.
Pesaran, M.H. (1997), The Role of Economic Theory in Modelling the Long Run, Economic Journal 107, 178–91.
Pesaran M.H. (2004), General Diagnostic Tests for Cross Section Dependence in Panels, *Cambridge Working Papers in Economics, No. 435*, University of Cambridge, and CESifo Working Paper Series No. 1229.
Pesaran, M.H., (2006), Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure, Econometrica, 74, 967–1012.
Pesaran, M.H. (2007), A Simple Panel Unit Root Test in the Presence of Cross Section Dependence, Journal of Applied Econometrics, 22, 265–312.
Pesaran, M.H. and Y. Shin (2002), Long-Run Structural Modelling, Econometric Reviews, 21, 49–87.
Pesaran, M.H., T. Schuermann, and S.M. Weiner (2004), Modelling Regional Interdependencies Using a Global Error-Correcting Macroeconometric Model, Journal of Business and Economic Statistics, 22, 129–162.
Pesaran, M.H., Y. Shin and R.P. Smith (1999), Pooled Mean Group Estimation of Dynamic Heterogeneous Panels, Journal of the American Statistical Association, 94, 621–624.
Pesaran, M.H., Y. Shin and R.J. Smith (2000), Structural Analysis of Vector Error Correction Models with Exogenous I(1) Variables, Journal of Econometrics, 97, 293–343.
Pesaran, M.H., R. Smith, and K.S. Im (1996), Dynamic linear models for Heterogenous Panels. In: Matyas, L., Sevestre, P. (eds), The Econometrics of Panel Data: A Handbook of the Theory with Applications, second revised edition, 145–195.
Pesaran, M.H. and Tosetti, E. (2007), Large Panels with Common Factors and Spatial Correlations, Faculty of Economics and Politics, University of Cambridge.
Pesaran, M.H. and T. Yamagata (2008), Testing Slope Homogeneity in Large Panels, Journal of Econometrics, 142, 50–93.
Phillips, P.C.B. (1991), Optimal Inference in Co-integrated Systems, Econometrica, 59, 282–306.
Phillips, P.C.B. (1994), Some Exact Distribution Theory for Maximum Likelihood Estimators of Cointegrating Coefficients in Error Correction Models, Econometrica, 62, 73–93.
Phillips, P.C.B. and B.E. Hansen (1990), Statistical Inference in Instrumental Variable Regression with I(1) Processes, Review of Economic Studies 57: 99–125.
Phillips, P.C.B. and H.R. Moon (1999), Linear Regression Limit Theory for Nonstationary Panel Data, Econometrica, 67, 1057–1111.
Phillips, P.C.B. and S. Ouliaris (1990), Asymptotic Properties of Residual Based Tests for Cointegration, Econometrica, 58, 165–193.
Phillips, P.C.B. and D. Sul (2003a), Dynamic Panel Estimation and Homogeneity Testing Under Cross Section Dependence, Econometrics Journal, 6, 217–259.
Phillips, P.C.B. and D. Sul (2003b), The Elusive Empirical Shadow of Growth Convergence, Cowles Foundation Discussion Paper 98, Yale University.
Ploberger, W. and P.C.B. Phillips (2002), Optimal Testing for Unit Roots in Panel Data, unpublished.
Robinson, P.M. (1994), Efficient Tests of Nonstationary Hypotheses, Journal of the American Statistical Association, 89, 1420–1437.
Saikkonen, P. (1991), Asymptotic Efficient Estimation of Cointegration Regressions, Econometric Theory, 7, 1–21.
Schmidt, P. and P.C.B. Phillips (1992), LM Test for a Unit Root in the Presence of Deterministic Trends, Oxford Bulletin of Economics and Statistics 54, 257–287.
Shin, Y. (1994), A Residual-based Test of the Null of Cointegration Against the Alternative of no Cointegration, Econometric Theory, 10, 91–115.
Smith, V., S. Leybourne, T.-H. Kim and P. Newbold (2004), More Powerful Panel Data Unit Root Tests With an Application to Mean Reversion in Real Exchange Rates, Journal of Applied Econometrics, 19, 147–170.
Tanaka, K. (1990), Testing for a Moving Average Root, Econometric Theory, 6, 433–444.

Tzavalis, E. (2002), Structural Breaks and Unit Root Tests for Short Panels, Queen Mary University of London, unpublished.
Wagner, M. (2007), On PPP, Unit Roots and Panels, forthcoming in: Empirical Economics.
Westerlund, J. (2005a), New Simple Tests for Panel Cointegration', Econometric Reviews, 24, 3, 297–316.
Westerlund, J. (2005b), A Panel CUSUM Test of the Null of Cointegration, Oxford Bulletin of Economics and Statistics, 62, 231–262.
Westerlund, J. (2005c), Data Dependent Endogeneity Correction in Cointegrated Panels, Oxford Bulletin of Economics and Statistics, 67, 691–705.
Westerlund, J. (2007a), A Note on the Pooling of Individual PANIC Unit Root Tests, Luund University, unpublished.
Westerlund, J. (2007b), Estimating Cointegrated Panels with Common Factors and The Forward Rate Unbiasedness Hypothesis, Journal of Financial Econometrics, 3, 491–522.
Westerlund, J. (2008), Some Cautions on the LLC Panel Unit Root Test, forthcoming in: Empirical Economics.

Chapter 10
Measurement Errors and Simultaneity

Erik Biørn and Jayalakshmi Krishnakumar

10.1 Introduction

This chapter is concerned with the problem of endogeneity of certain explanatory variables in a regression equation. There are two potential sources of endogeneity in a panel data model with individual and time specific effects : (i) correlation between explanatory variables and specific effects (when treated random) and (ii) correlation between explanatory variables and the residual/idiosyncratic error term.

The first case was extensively dealt with in Chap. 4 of this book and hence we will not go into it here. In this chapter we are more concerned with a non-zero correlation between the explanatory variables and the overall error consisting of both the specific effect and the genuine disturbance term. One might call it double endogeneity as opposed to the single endogeneity in the former situation.

In this chapter we consider two major causes of this double endogeneity encountered in practical situations. One of them is the presence of measurement errors in the explanatory variables. This will be the object of study of Sect. 10.2. Another major source is the simultaneity problem that arises when the regression equation is one of several structural equations of a simultaneous model and hence contains current endogenous explanatory variables. Sect. 10.3 will look into this problem in detail and Sect. 10.4 concludes the chapter.

10.2 Measurement Errors and Panel Data

A familiar and notable property of the Ordinary Least Squares (OLS) when there are random measurement errors (errors-in-variables, EIV) in the regressors is that the coefficient estimators are inconsistent. In the one regressor case (or the multiple

Erik Biørn
Department of Economics, University of Oslo, P.O. Box 1095, Blindern, 0317 Oslo, Norway, e-mail: erik.biorn@econ.uio.no

Jayalakshmi Krishnakumar
Department of Econometrics, University of Geneva, 40 Bd. Du pont d'Ave, CH-1211 Geneva 4, Switzerland, e-mail: jaya.krishnakumar@metri.unige.ch

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

regressor case with uncorrelated regressors) under standard assumptions, the slope coefficient estimator is biased towards zero, often denoted as *attenuation*. More seriously, unless some 'extraneous' information is available, e.g. the existence of valid parameter restrictions or valid instruments for the error-ridden regressors, slope coefficients cannot (in general) be identified from standard data [see Fuller (1987, Sect. 1.1.3)].[1] This *lack of identification* in EIV models, however, relates to *uni-dimensional* data, i.e., pure (single or repeated) cross-sections or pure time-series. If the variables are observed as panel data, exhibiting *two-dimensional* variation, it may be possible to handle the EIV identification problem and estimate slope coefficients consistently without extraneous information, provided that the distribution of the latent regressors and the measurement errors satisfy certain weak conditions.

Briefly, the reason why the existence of variables observed along two dimensions makes the EIV identification problem easier to solve, is partly (i) the *repeated measurement* property of panel data, so that the measurement error problem can be reduced by taking averages, which, in turn, may show sufficient variation to permit consistent estimation, and partly (ii) the larger set of *other linear data transformations* available for estimation. Such transformations, involving several individuals or several periods, may be needed to take account of uni-dimensional 'nuisance variables' like unobserved individual or period specific *heterogeneity*, which are potentially correlated with the regressor.

Our focus is on the estimation of linear, static regression equations from balanced panel data with additive, random measurement errors in the regressors by means of methods utilizing instrumental variables (IVs). The panel data available to an econometrician are frequently from individuals, firms, or other kinds of micro units, where not only observation errors in the narrow sense, but also departures between theoretical variable definitions and their observable counterparts in a wider sense may be present.

From the panel data literature which disregards the EIV problem we know that the effect of, for example, additive (fixed or random) individual heterogeneity within a linear model can be eliminated by deducting individual means, taking differences over periods, etc. [see Baltagi (2001, Chap. 2) and Hsiao (2003, Sect. 1.1)]. Such transformations, however, may magnify the variation in the measurement error component of the observations relative to the variation in the true structural component, i.e., they may increase the 'noise/signal ratio'. Hence, data transformations intended to 'solve' the unobserved heterogeneity problem in estimating slope coefficients may aggravate the EIV problem. Several familiar estimators for panel data models, including the fixed effects within-group and between-group estimators, and the random effects Generalised Least Squares (GLS) estimators will then be inconsistent, the bias depending, inter alia, on the way in which the number of individuals and/or

[1] Identification under non-normality of the true regressor is possible, by utilizing moments of the distribution of the observable variables of order higher than the second [see Reiersøl (1950)]. Even under non-identification, bounds on the parameters can be established from the distribution of the observable variables [see Fuller (1987, p. 11)]. These bounds may be wide or narrow, depending on the covariance structure of the variables; see Klepper and Leamer (1984), Bekker et al. (1987), and Erickson (1993).

periods tend to infinity and on the heterogeneity of the measurement error process; see Griliches and Hausman (1986) and Biørn (1992, 1996). Such inconsistency problems will not be dealt with here. Neither will we consider the idea of constructing consistent estimators by combining two or more inconsistent ones with different probability limits. Several examples are given in Griliches and Hausman (1986), Biørn (1996), and Wansbeek and Meijer (2000, Sect. 6.9).

The procedures to be considered in this section have two basic characteristics: First, *a mixture of level and difference variables* are involved. Second, the orthogonality conditions derived from the EIV structure – involving levels and differences over one or more than one periods – are not all essential, some are redundant. Our estimation procedures are of two kinds: (A) Transform the equation to *differences* and estimate it by IV or GMM, using as IVs *level* values of the regressors and/or regressands for other periods. (B) Keep the equation in *level* form and estimate it by IV or GMM, using as IVs *differenced* values of the regressors and/or regressands for other periods. In both cases, the differencing serves to eliminate individual heterogeneity, which is a potential nuisance since it may be correlated with the latent regressor vector. These procedures resemble, to some extent, procedures for autoregressive (AR) models for panel data without measurement errors (mostly AR(1) equations with individual heterogeneity and often with exogenous regressors added) discussed, inter alia, by Anderson and Hsiao (1981, 1982), Holtz-Eakin et al. (1988), Arellano and Bond (1991), Arellano and Bover (1995), Ahn and Schmidt (1995), Blundell and Bond (1998), and Harris et al. (2007).

If the distribution of the latent regressor vector is *not* time invariant and the second order moments of the measurement errors and disturbances are structured to some extent, a large number of consistent IV estimators of the coefficient of the latent regressor vector exist. Their consistency is robust to potential correlation between the individual heterogeneity and the latent regressor. Serial correlation or non-stationarity of the latent regressor is favourable from the point of view of identification and estimability of the coefficient vector.

The literature dealing specifically with panel data with measurement errors is not large. The (A) procedures above extend and modify procedures described in Griliches and Hausman (1986), which is the seminal article on measurement errors in panel data, at least in econometrics. Extensions are discussed in Wansbeek and Koning (1991), Biørn (1992, 1996, 2000, 2003), and Biørn and Klette (1998, 1999), and Wansbeek (2001). Paterno et al. (1996) consider Maximum Likelihood analysis of panel data with measurement errors and is not related to the (A) and (B) procedures to be discussed here.

10.2.1 Model and Orthogonality Conditions

Consider a panel data set with $N\ (\geq 2)$ individuals observed in $T\ (\geq 2)$ periods and a relationship between y (observable scalar) and a $(K \times 1)$ vector $\boldsymbol{\xi}$ (latent),

$$y_{it} = c + \boldsymbol{\xi}'_{it}\boldsymbol{\beta} + \alpha_i + u_{it}, \qquad i = 1,\ldots,N;\ t = 1,\ldots,T, \tag{10.1}$$

where $(y_{it}, \boldsymbol{\xi}_{it})$ is the value of $(y, \boldsymbol{\xi})$ for individual i in period t, c is a scalar constant, $\boldsymbol{\beta}$ is a $(K \times 1)$ vector and α_i is a zero (marginal) mean individual effect, which we consider as random and potentially correlated with $\boldsymbol{\xi}_{it}$, and u_{it} is a zero mean disturbance, which may also contain a measurement error in y_{it}. We observe

$$\boldsymbol{x}_{it} = \boldsymbol{\xi}_{it} + \boldsymbol{v}_{it}, \qquad i = 1,\ldots,N;\ t = 1,\ldots,T, \tag{10.2}$$

where $\boldsymbol{v}_{it}$ is a zero mean $(K \times 1)$ vector of measurement errors. Hence,

$$y_{it} = c + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \quad \varepsilon_{it} = \alpha_i + u_{it} - \boldsymbol{v}'_{it}\boldsymbol{\beta}. \tag{10.3}$$

We can eliminate α_i from (10.3) by taking arbitrary backward differences $\Delta y_{it\theta} = y_{it} - y_{i\theta}$, $\Delta \boldsymbol{x}_{it\theta} = \boldsymbol{x}_{it} - \boldsymbol{x}_{i\theta}$, etc., giving

$$\Delta y_{it\theta} = \Delta \boldsymbol{x}'_{it\theta}\boldsymbol{\beta} + \Delta\varepsilon_{it\theta}, \quad \Delta\varepsilon_{it\theta} = \Delta u_{it\theta} - \Delta \boldsymbol{v}'_{it\theta}\boldsymbol{\beta}. \tag{10.4}$$

We assume that $(\boldsymbol{\xi}_{it}, u_{it}, \boldsymbol{v}_{it}, \alpha_i)$ are independent across individuals [which excludes random period specific components in $(\boldsymbol{\xi}_{it}, u_{it}, \boldsymbol{v}_{it})$], and make the following *basic orthogonality assumptions*:

Assumption (A):
$$\begin{aligned} &\mathsf{E}(\boldsymbol{v}_{it}u_{i\theta}) = \mathsf{E}(\boldsymbol{\xi}_{it}u_{i\theta}) = \mathsf{E}(\alpha_i\boldsymbol{v}_{it}) = \mathbf{0}_{K1}, \\ &\mathsf{E}(\boldsymbol{\xi}_{i\theta} \otimes \boldsymbol{v}'_{it}) = \mathbf{0}_{KK}, \\ &\mathsf{E}(\alpha_i u_{it}) = 0, \end{aligned} \qquad \begin{aligned} &i = 1,\ldots,N, \\ &t,\theta = 1,\ldots,T, \end{aligned}$$

where $\mathbf{0}_{mn}$ denotes the $(m \times n)$ zero matrix and $\otimes$ is the Kronecker product operator. Regarding the temporal structure of the measurement errors and disturbances, we assume either that

$$\begin{aligned} &\text{Assumption (B1):} && \mathsf{E}(\boldsymbol{v}_{it}\boldsymbol{v}'_{i\theta}) = \mathbf{0}_{KK}, && |t-\theta| > \tau, \\ &\text{Assumption (C1):} && \mathsf{E}(u_{it}u_{i\theta}) = 0, && |t-\theta| > \tau, \end{aligned}$$

where τ is a non-negative integer, indicating the order of the serial correlation, or

$$\begin{aligned} &\text{Assumption (B2):} && \mathsf{E}(\boldsymbol{v}_{it}\boldsymbol{v}'_{i\theta}) \text{ is invariant to } t,\theta, && t \neq \theta, \\ &\text{Assumption (C2):} && \mathsf{E}(u_{it}u_{i\theta}) \text{ is invariant to } t,\theta, && t \neq \theta, \end{aligned}$$

which allows for time invariance of the autocovariances. The latter will, for instance, be satisfied if the measurement errors and the disturbances have individual specific components, say $\boldsymbol{v}_{it} = \boldsymbol{v}_{1i} + \boldsymbol{v}_{2it}$, $u_{it} = u_{1i} + u_{2it}$, where $\boldsymbol{v}_{1i}$, $\boldsymbol{v}_{2it}$, u_{1i}, and u_{2it} are independent IID processes.

The final set of assumptions relate to the *distribution of the latent regressor vector* $\boldsymbol{\xi}_{it}$:

Assumption (D1):	$\mathsf{E}(\boldsymbol{\xi}_{it})$	is invariant to t,
Assumption (D2):	$\mathsf{E}(\alpha_i \boldsymbol{\xi}_{it})$	is invariant to t,
Assumption (E):	$\operatorname{rank}(\mathsf{E}[\boldsymbol{\xi}_{ip}(\Delta\boldsymbol{\xi}'_{it\theta})]) = K$	for some p,t,θ different,

Assumptions (D1) and (D2) hold when $\boldsymbol{\xi}_{it}$ is stationary for all i [(D1) alone imposing mean stationarity]. Assumption (E) *imposes non-*IID *and some form of autocorrelation or non-stationarity on* $\boldsymbol{\xi}_{it}$. It excludes, for example, the case where $\boldsymbol{\xi}_{it}$ has an individual specific component, so that $\boldsymbol{\xi}_{it} = \boldsymbol{\xi}_{1i} + \boldsymbol{\xi}_{2it}$, where $\boldsymbol{\xi}_{1i}$ and $\boldsymbol{\xi}_{2it}$ are independent (vector) IID processes.

Assumptions (A)–(E) do not impose much structure on the first and second order moments of the u_{it}s, $\boldsymbol{v}_{it}$s, $\boldsymbol{\xi}_{it}$s and α_is. This has both its pros and cons. It is possible to structure this distribution more strongly, for instance assuming homoskedasticity and normality of u_{it}, v_{it}, and α_i, and normality of $\boldsymbol{\xi}_{it}$. Exploiting this stronger structure, e.g., by taking a LISREL of LIML approach, we might obtain more efficient (but potentially less robust) estimators by operating on the full covariance matrix of the y_{it}s and the $\boldsymbol{x}_{it}$s rather than eliminating the α_is by differencing. Other extensions are elaborated in Sect. 10.2.7.

10.2.2 Identification and the Structure of the Second Order Moments

The distribution of $(\boldsymbol{\xi}_{it}, u_{it}, \boldsymbol{v}_{it}, \alpha_i)$ must satisfy some conditions to make identification of $\boldsymbol{\beta}$ possible. The nature of these conditions can be illustrated as follows. Assume, for simplicity, that this distribution is the same for all individuals and that (A) holds, and let

$$\begin{aligned} &\mathsf{C}(\boldsymbol{\xi}_{it}, \boldsymbol{\xi}_{i\theta}) = \boldsymbol{\Sigma}_{t\theta}^{\xi\xi}, \quad \mathsf{E}(\boldsymbol{\xi}_{it}\alpha_i) = \boldsymbol{\Sigma}_t^{\xi\alpha}, \quad \mathsf{E}(\alpha_i^2) = \sigma^{\alpha\alpha}, && i = 1,\ldots,N, \\ &\mathsf{E}(\boldsymbol{v}_{it}\boldsymbol{v}'_{i\theta}) = \boldsymbol{\Sigma}_{t\theta}^{vv}, \quad \mathsf{E}(u_{it}u_{i\theta}) = \sigma_{t\theta}^{uu}, && t,\theta = 1,\ldots,T, \end{aligned}$$

where C denotes the covariance matrix operator. It then follows from (10.1) and (10.2) that the second order moments of the observable variables can be expressed as

$$\begin{aligned} \mathsf{C}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i\theta}) &= \boldsymbol{\Sigma}_{t\theta}^{\xi\xi} + \boldsymbol{\Sigma}_{t\theta}^{vv}, \\ \mathsf{C}(\boldsymbol{x}_{it}, y_{i\theta}) &= \boldsymbol{\Sigma}_{t\theta}^{\xi\xi}\boldsymbol{\beta} + \boldsymbol{\Sigma}_t^{\xi\alpha}, \\ \mathsf{C}(y_{it}, y_{i\theta}) &= \boldsymbol{\beta}'\boldsymbol{\Sigma}_{t\theta}^{\xi\xi}\boldsymbol{\beta} + (\boldsymbol{\Sigma}_t^{\xi\alpha})'\boldsymbol{\beta} + \beta'\boldsymbol{\Sigma}_\theta^{\xi\alpha} + \sigma_{t\theta}^{uu} + \sigma^{\alpha\alpha}, \end{aligned} \qquad \begin{matrix} i = 1,\ldots,N, \\ t,\theta = 1,\ldots,T. \end{matrix} \tag{10.5}$$

The identifiability of $\boldsymbol{\beta}$ from second order moments in general depends on whether or not knowledge of $\mathsf{C}(\boldsymbol{x}_{it}, \boldsymbol{x}_{i\theta})$, $\mathsf{C}(\boldsymbol{x}_{it}, y_{i\theta})$, and $\mathsf{C}(y_{it}, y_{i\theta})$ for all available t and θ is sufficient for obtaining a unique solution for $\boldsymbol{\beta}$ from (10.5), given the restrictions imposed on the $\boldsymbol{\Sigma}_{t\theta}^{\xi\xi}$s, $\boldsymbol{\Sigma}_t^{\xi\alpha}$s, $\sigma_{t\theta}^{uu}$s, and $\sigma^{\alpha\alpha}$. The answer, in general, depends on T and K. With no further information, the number of elements in

$\mathsf{C}(\boldsymbol{x}_{it},\boldsymbol{x}_{i\theta})$, and $\mathsf{C}(y_{it},y_{i\theta})$ (all of which can be estimated consistently from corresponding sample moments under weak conditions) equal the number of unknown elements in $\boldsymbol{\Sigma}^{vv}_{t\theta}$ and $\sigma^{uu}_{t\theta}$, which is $\frac{1}{2}KT(KT+1)$ and $\frac{1}{2}T(T+1)$, respectively. Then $\sigma^{\alpha\alpha}$ cannot be identified, and $\mathsf{C}(\boldsymbol{x}_{it},y_{i\theta})$ contains the only additional information available for identifying $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}^{\xi\xi}_{t\theta}$, and $\boldsymbol{\Sigma}^{\xi\alpha}_{t}$, given the restrictions imposed on the latter two matrices.

Consider two extreme cases. First, if $T=1$, i.e., if we only have cross-section data, and no additional restrictions are imposed, there is an identification problem for any K. Second, if $T>2$ and $\boldsymbol{\xi}_{it}\sim\mathsf{IID}(\boldsymbol{\mu}^{\xi},\boldsymbol{\Sigma}^{\xi\xi})$, $\boldsymbol{v}_{it}\sim\mathsf{IID}(\boldsymbol{0}_{1,K},\boldsymbol{\Sigma}^{vv})$, $u_{it}\sim\mathsf{IID}(0,\sigma^{uu})$, $\alpha_i\sim\mathsf{IID}(0,\sigma^{\alpha\alpha})$, we also have lack of identification in general. We get an essentially similar conclusion when the autocovariances of $\boldsymbol{\xi}_{it}$ are time invariant and it is IID across i. From (10.5) we then get

$$\begin{aligned}
&\mathsf{C}(\boldsymbol{x}_{it},\boldsymbol{x}_{i\theta})=\delta_{t\theta}(\boldsymbol{\Sigma}^{\xi\xi}+\boldsymbol{\Sigma}^{vv}),\\
&\mathsf{C}(\boldsymbol{x}_{it},y_{i\theta})=\delta_{t\theta}\boldsymbol{\Sigma}^{\xi\xi}\boldsymbol{\beta},\\
&\mathsf{C}(y_{it},y_{i\theta})=\delta_{t\theta}(\boldsymbol{\beta}'\boldsymbol{\Sigma}^{\xi\xi}\boldsymbol{\beta}+\sigma^{uu})+\sigma^{\alpha\alpha},
\end{aligned}\tag{10.6}$$

where $\delta_{t\theta}=1$ for $t=\theta$ and $=0$ for $t\neq\theta$, and so we are essentially in the same situation with regard to identifiability of $\boldsymbol{\beta}$ as when $T=1$. The 'cross-period' equations ($t\neq\theta$) then serve no other purpose than identification of $\sigma^{\alpha\alpha}$, and whether $T=1$ or $T>1$ realizations of $\mathsf{C}(\boldsymbol{x}_{it},\boldsymbol{x}_{it})$, $\mathsf{C}(\boldsymbol{x}_{it},y_{it})$, and $\mathsf{C}(y_{it},y_{it})$ are available in (10.6) is immaterial to the identifiability of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}^{\xi\xi}$, $\boldsymbol{\Sigma}^{vv}$, and σ^{uu}. In intermediate situations, identification may be ensured when $T\geq 2$. These examples illustrate that in order to ensure identification of the slope coefficient vector from panel data, there should not be 'too much structure' on the second order moments of the latent exogenous regressors along the time dimension, and not 'too little structure' on the second order moments of the errors and disturbances along the time dimension.

10.2.3 Moment Conditions

A substantial number of (linear and non-linear) moment conditions involving y_{it}, $\boldsymbol{x}_{it}$, and ε_{it} can be derived from Assumptions (A)–(E). Since (10.1)–(10.3) and Assumption (A) imply

$$\begin{aligned}
\mathsf{E}(\boldsymbol{x}_{it}\boldsymbol{x}'_{i\theta})&=\mathsf{E}(\boldsymbol{\xi}_{it}\boldsymbol{\xi}'_{i\theta})+\mathsf{E}(\boldsymbol{v}_{it}\boldsymbol{v}'_{i\theta}),\\
\mathsf{E}(\boldsymbol{x}_{it}y_{i\theta})&=\mathsf{E}(\boldsymbol{\xi}_{it}\boldsymbol{\xi}'_{i\theta})\boldsymbol{\beta}+\mathsf{E}[\boldsymbol{\xi}_{it}(\alpha_i+c)],\\
\mathsf{E}(y_{it}y_{i\theta})&=c^2+\mathsf{E}(\alpha_i^2)+\boldsymbol{\beta}'\mathsf{E}(\boldsymbol{\xi}_{it}\boldsymbol{\xi}'_{i\theta})\boldsymbol{\beta}+\boldsymbol{\beta}'\mathsf{E}[\boldsymbol{\xi}_{it}(\alpha_i+c)]\\
&\quad+\mathsf{E}[(\alpha_i+c)\boldsymbol{\xi}'_{i\theta}]\boldsymbol{\beta}+\mathsf{E}(u_{it}u_{i\theta}),\\
\mathsf{E}(\boldsymbol{x}_{it}\varepsilon_{i\theta})&=\mathsf{E}(\boldsymbol{\xi}_{it}\alpha_i)-\mathsf{E}(\boldsymbol{v}_{it}\boldsymbol{v}'_{i\theta})\boldsymbol{\beta},\\
\mathsf{E}(y_{it}\varepsilon_{i\theta})&=\boldsymbol{\beta}'\mathsf{E}(\boldsymbol{\xi}_{it}\alpha_i)+\mathsf{E}(\alpha_i^2)+\mathsf{E}(u_{it}u_{i\theta}),
\end{aligned}$$

we can derive moment equations involving observable variables in levels and differences:

$$\mathsf{E}[\boldsymbol{x}_{ip}(\Delta\boldsymbol{x}'_{it\theta})] = \mathsf{E}[\boldsymbol{\xi}_{ip}(\Delta\boldsymbol{\xi}'_{it\theta})] + \mathsf{E}[\boldsymbol{v}_{ip}(\Delta\boldsymbol{v}'_{it\theta})], \tag{10.7}$$

$$\mathsf{E}[\boldsymbol{x}_{ip}(\Delta y'_{it\theta})] = \mathsf{E}[\boldsymbol{\xi}_{ip}(\Delta\boldsymbol{\xi}'_{it\theta})]\boldsymbol{\beta}, \tag{10.8}$$

$$\mathsf{E}[(\Delta\boldsymbol{x}_{ipq})y_{it}] = \mathsf{E}[(\Delta\boldsymbol{\xi}_{ipq})\boldsymbol{\xi}'_{it}]\boldsymbol{\beta} + \mathsf{E}[(\Delta\boldsymbol{\xi}_{ipq})(\alpha_i + c)], \tag{10.9}$$

as well as moment equations involving observable variables and errors/disturbances:

$$\mathsf{E}[\boldsymbol{x}_{ip}(\Delta\varepsilon_{it\theta})] = -\mathsf{E}[\boldsymbol{v}_{ip}(\Delta\boldsymbol{v}'_{it\theta})]\boldsymbol{\beta}, \tag{10.10}$$

$$\mathsf{E}[y_{ip}(\Delta\varepsilon_{it\theta})] = \mathsf{E}[u_{ip}(\Delta u_{it\theta})], \tag{10.11}$$

$$\mathsf{E}[(\Delta\boldsymbol{x}_{ipq})\varepsilon_{it}] = \mathsf{E}[(\Delta\boldsymbol{\xi}_{ipq})\alpha_i] - \mathsf{E}[(\Delta\boldsymbol{v}_{ipq})\boldsymbol{v}'_{it}]\boldsymbol{\beta}, \tag{10.12}$$

$$\mathsf{E}[(\Delta y_{ipq})\varepsilon_{it}] = \boldsymbol{\beta}'\mathsf{E}[(\Delta\boldsymbol{\xi}_{ipq})\alpha_i] + \mathsf{E}[(\Delta u_{ipq})u_{it}], \quad t,\theta,p,q = 1,\ldots,T. \tag{10.13}$$

Not all of the equations in (10.7)–(10.13), whose number is substantial even for small T, are, of course, independent. Depending on which (B), (C), and (D) assumptions are valid, some terms on the right hand side of (10.9)–(10.13) will vanish. Precisely, if $T > 2$, then (10.3), (10.5), and (10.10)–(10.13) imply the following moment conditions, or *orthogonality conditions (OC)*, on the observable variables and the errors and disturbances

$$\begin{aligned}&\text{(B2), or (B1) with } |t-p|, |\theta-p| > \tau, t \neq \theta\\ &\quad\Longrightarrow \mathsf{E}[\boldsymbol{x}_{ip}(\Delta\varepsilon_{it\theta})] = \mathsf{E}[\boldsymbol{x}_{ip}(\Delta y_{it\theta})] - \mathsf{E}[\boldsymbol{x}_{ip}(\Delta\boldsymbol{x}'_{it\theta})]\boldsymbol{\beta} = \mathbf{0}_{K1}.\end{aligned} \tag{10.14}$$

$$\begin{aligned}&\text{(C2), or (C1) with } |t-p|, |\theta-p| > \tau, t \neq \theta\\ &\quad\Longrightarrow \mathsf{E}[y_{ip}(\Delta\varepsilon_{it\theta})] = \mathsf{E}[y_{ip}(\Delta y_{it\theta})] - \mathsf{E}[y_{ip}(\Delta\boldsymbol{x}'_{it\theta})]\boldsymbol{\beta} = 0.\end{aligned} \tag{10.15}$$

$$\begin{aligned}&\text{(D1), (D2) and (B2), or (B1) with } |t-p|, |t-q| > \tau, p \neq q\\ &\quad\Longrightarrow \mathsf{E}[(\Delta\boldsymbol{x}_{ipq})\varepsilon_{it}] = \mathsf{E}[(\Delta\boldsymbol{x}_{ipq})y_{it}] - \mathsf{E}[(\Delta\boldsymbol{x}_{ipq})\boldsymbol{x}_{it}]\boldsymbol{\beta} = \mathbf{0}_{K1}.\end{aligned} \tag{10.16}$$

$$\begin{aligned}&\text{(D1), (D2), and (C2), or (C1) with } |t-p|, |t-q| > \tau, p \neq q\\ &\quad\Longrightarrow \mathsf{E}[(\Delta y_{ipq})\varepsilon_{it}] = \mathsf{E}[(\Delta y_{ipq})y_{it}] - \mathsf{E}[(\Delta y_{ipq})\boldsymbol{x}'_{it}]\boldsymbol{\beta} = 0.\end{aligned} \tag{10.17}$$

The treatment of the intercept term c in constructing (10.16) and (10.17) needs a comment. When the mean stationarity assumption (D1) holds, using IVs in differences annihilates c in the moment equations, since then $\mathsf{E}(\Delta\boldsymbol{x}_{ipq}) = \mathbf{0}_{K1}$ and $\mathsf{E}(\Delta y_{ipq}) = 0$. If, however, we *relax* (D1), which is unlikely to hold in many practical situations, we get

$$\begin{aligned}\mathsf{E}[(\Delta\boldsymbol{x}_{ipq})\varepsilon_{it}] &= \mathsf{E}[(\Delta\boldsymbol{x}_{ipq})y_{it}] - \mathsf{E}[\Delta\boldsymbol{x}_{ipq}]c - \mathsf{E}[(\Delta\boldsymbol{x}_{ipq})\boldsymbol{x}'_{it}]\boldsymbol{\beta} = \mathbf{0}_{K1},\\ \mathsf{E}[(\Delta y_{ipq})\varepsilon_{it}] &= \mathsf{E}[(\Delta y_{ipq})y_{it}] - \mathsf{E}[\Delta y_{ipq}]c - \mathsf{E}[(\Delta y_{ipq})\boldsymbol{x}'_{it}]\boldsymbol{\beta} = 0.\end{aligned}$$

Using $\mathsf{E}(\varepsilon_{it}) = \mathsf{E}(y_{it}) - c - E(\boldsymbol{x}'_{it})\boldsymbol{\beta} = 0$ to eliminate c leads to the following modifications of (10.16) and (10.17):

(D1), (D2) and (B2), or (B1) with $|t-p|,|t-q|>\tau, p\neq q$,
$\Longrightarrow \mathsf{E}[(\Delta \boldsymbol{x}_{ipq})\varepsilon_{it}] = \mathsf{E}[(\Delta \boldsymbol{x}_{ipq})(y_{it}-\mathsf{E}(y_{it}))] - \mathsf{E}[(\Delta \boldsymbol{x}_{ipq})(\boldsymbol{x}'_{it}-\mathsf{E}(\boldsymbol{x}'_{it}))]\boldsymbol{\beta} = \mathbf{0}_{K1}$.

(D1), (D2), and (C2), or (C1) with $|t-p|,|t-q|>\tau, p\neq q$,
$\Longrightarrow \mathsf{E}[(\Delta y_{ipq})\varepsilon_{it}] = \mathsf{E}[(\Delta y_{ipq})(y_{it}-\mathsf{E}(y_{it}))] - \mathsf{E}[(\Delta y_{ipq})(\boldsymbol{x}'_{it}-\mathsf{E}(\boldsymbol{x}'_{it}))]\boldsymbol{\beta} = 0$.

To implement these modified OCs in the GMM procedures to be described below for the level equation, we could replace $\mathsf{E}(y_{it})$ and $\mathsf{E}(\boldsymbol{x}_{it})$ by corresponding global or period specific sample means.

The conditions in (10.14)–(10.17) are not all independent. Some are redundant, since they can be derived as linear combinations of other conditions.[2] We confine attention to (10.14) and (10.16), since (10.15) and (10.17) can be treated similarly. When $\tau = 0$, the total number of OCs in both (10.14) and (10.16) is $\frac{1}{2}KT(T-1)(T-2)$. Below, we prove that

(a) When (B2) and (C2), or (B1) and (C1) with $\tau = 0$, are satisfied, all OCs in (10.14) can be constructed from *all admissible OCs relating to equations differenced over one period and a subset of OCs relating to differences over two periods.* When (B1) and (C1) are satisfied with an arbitrary τ, all OCs in (10.14) can be constructed from *all admissible OCs relating to equations differenced over one period and a subset of OCs relating to differences over* $2(\tau+1)$ *periods*.

(b) When (B2) and (C2), or (B1) and (C1) with $\tau = 0$, are satisfied all OCs in (10.16) can be constructed from *all admissible OCs relating to IVs differenced over one period and a subset of IVs differenced over two periods*. When (B1) and (C1) are satisfied with an arbitrary τ, all OCs in (10.16) can be constructed from *all admissible OCs relating to IVs differenced over one period and a subset of IVs differenced over* $2(\tau+1)$ *periods*.

We denote the non-redundant conditions defined by (a) and (b) as *essential* OCs. Since (10.14) and (10.16) are symmetric, we prove only (a) and derive (b) by way of analogy.

Since $\boldsymbol{x}_{ip}\Delta\varepsilon_{it\theta} = \boldsymbol{x}_{ip}(\sum_{j=\theta+1}^{t}\Delta\varepsilon_{ij,j-1})$, we see that if (hypothetically) all $p = 1,\ldots,T$ combined with all $t > \theta$ would have given admissible OCs, (10.14) for differences over $2,3,\ldots,T-1$ periods could have been constructed from the conditions relating to one-period differences only. However, since $(t,\theta) = (p,p-1),(p+1,p)$ are inadmissible, and [when (B2) holds] $(t,\theta) = (p+1,p-1)$ is admissible, we have to distinguish between the cases where p is strictly outside and strictly inside the interval (θ,t). From the identities

$$\boldsymbol{x}_{ip}\Delta\varepsilon_{it\theta} = \boldsymbol{x}_{ip}(\textstyle\sum_{j=\theta+1}^{t}\Delta\varepsilon_{ij,j-1}) \text{ for } p = 1,\ldots,\theta-1,\ t+1,\ldots,T,$$
$$\boldsymbol{x}_{ip}\Delta\varepsilon_{it\theta} = \boldsymbol{x}_{ip}(\textstyle\sum_{j=\theta+1}^{p-1}\Delta\varepsilon_{ij,j-1} + \Delta\varepsilon_{i,p+1,p-1} + \sum_{j=p+2}^{t}\Delta\varepsilon_{ij,j-1}) \text{ for } p = \theta+1,\ldots,t-1,$$

when taking expectations, we then obtain

[2] This redundancy problem is discussed in Biørn (2000). Essential and redundant moment conditions in AR models for panel data are discussed in Ahn and Schmidt (1995), Arellano and Bover (1995), and Blundell and Bond (1998). A general treatment of redundancy of moment conditions in GMM estimation is found in Breusch et al. (1999).

Proposition 10.1.
A. When (B2) and (C2) are satisfied, then

(a) $\mathsf{E}[\boldsymbol{x}_{ip}(\Delta\varepsilon_{it,t-1})] = \mathbf{0}_{K1}$ *for* $p = 1,\ldots,t-2,t+1,\ldots,T$; $t = 2,\ldots,T$ *are* $K(T-1)(T-2)$ *essential OCs for equations differenced over one period.*
(b) $\mathsf{E}[\boldsymbol{x}_{it}(\Delta\varepsilon_{it+1,t-1})] = \mathbf{0}_{K1}$ *for* $t = 2,\ldots,T-1$ *are* $K(T-2)$ *essential OCs for equations differenced over two periods.*
(c) *The other OCs are redundant: among the* $\frac{1}{2}KT(T-1)(T-2)$ *conditions in* (10.14)*, only a fraction* $2/(T-1)$*, are essential.*

B. When (B1) and (C1) are satisfied for an arbitrary τ, then

(a) $\mathsf{E}[\boldsymbol{x}_{ip}(\Delta\varepsilon_{it,t-1})] = \mathbf{0}_{K1}$ *for* $p = 1,\ldots,t-\tau-2,t+\tau+1,\ldots,T$; $t = 2,\ldots,T$ *are essential OCs for equations in one-period differences.*
(b) $\mathsf{E}[\boldsymbol{x}_{it}(\Delta\varepsilon_{it+\tau+1,t-\tau-1})] = \mathbf{0}_{K1}$ *for* $t = \tau+2,\ldots,T-\tau-1$ *are essential OCs for equations in* $2(\tau+1)$ *period differences.*
(c) *The other OCs in* (10.14) *are redundant.*

Symmetrically, from (10.16) we have

Proposition 10.2.
A. When (B2) and (C2) are satisfied, then

(a) $\mathsf{E}[(\Delta\boldsymbol{x}_{ip,p-1})\varepsilon_{it}] = \mathbf{0}_{K1}$ *for* $t = 1,\ldots,p-2,p+1,\ldots,T$; $p = 2,\ldots,T$ *are* $K(T-1)(T-2)$ *essential OCs for equations in levels, with IVs differenced over one period.*
(b) $\mathsf{E}[(\Delta\boldsymbol{x}_{it+1,t-1})\varepsilon_{it}] = \mathbf{0}_{K1}$ *for* $t = 2,\ldots,T-1$ *are* $K(T-2)$ *essential OCs for equations in levels, with IVs differenced over two periods.*
(c) *The other OCs are redundant: among the* $\frac{1}{2}KT(T-1)(T-2)$ *conditions in* (10.16)*, only a fraction* $2/(T-1)$*, are essential.*

B. When (B1) and (C1) are satisfied for an arbitrary τ, then

(a) $\mathsf{E}[(\Delta\boldsymbol{x}_{ip,p-1})\varepsilon_{it}] = \mathbf{0}_{K1}$ *for* $t = 1,\ldots,p-\tau-2,p+\tau+1,\ldots,T$; $p = 2,\ldots,T$ *are essential OCs for equations in levels, with IVs differenced over one period.*
(b) $\mathsf{E}[(\Delta\boldsymbol{x}_{it+\tau+1,t-\tau-1})\varepsilon_{it}] = \mathbf{0}_{K1}$ *for* $t = \tau+2,\ldots,T-\tau-1$ *are essential OCs for equations in levels, with IVs differenced over* $2(\tau+1)$ *periods.*
(c) *The other OCs in* (10.16) *are redundant.*

These propositions can be (trivially) modified to include also the essential and redundant OCs in the ys or the Δys, given in (10.15) and (10.17).

10.2.4 Estimators Constructed from Period Means

Several consistent estimators of $\boldsymbol{\beta}$ can be constructed from differenced period means. These estimators exploit the repeated measurement property of panel data, while the differencing removes the latent heterogeneity. From (10.3) we obtain

$$\Delta_s \bar{y}_{\cdot t} = \Delta_s \bar{\boldsymbol{x}}'_{\cdot t}\boldsymbol{\beta} + \Delta_s \bar{\varepsilon}_{\cdot t}, \qquad s = 1,\ldots,T-1; \quad t = s+1,\ldots,T, \tag{10.18}$$

$$(\bar{y}_{\cdot t} - \bar{y}) = (\bar{\boldsymbol{x}}_{\cdot t} - \bar{\boldsymbol{x}})'\boldsymbol{\beta} + (\bar{\varepsilon}_{\cdot t} - \bar{\varepsilon}), \qquad t = 1,\ldots,T, \tag{10.19}$$

where $\bar{y}_{\cdot t} = \frac{1}{N}\sum_i y_{it}$, $\bar{y} = \frac{1}{NT}\sum_i \sum_t y_{it}$, $\bar{\boldsymbol{x}}_{\cdot t} = \frac{1}{N}\sum_i \boldsymbol{x}_{it}$, $\bar{\boldsymbol{x}} = \frac{1}{NT}\sum_i \sum_t \boldsymbol{x}_{it}$, etc. and Δ_s denotes differencing over s periods. When (A) is satisfied, the (weak) law of large numbers implies, under weak conditions [confer McCabe and Tremayne (1993, Sect. 3.5)],[3] that $\operatorname{plim}(\bar{\varepsilon}_{\cdot t}) = 0$, $\operatorname{plim}(\bar{\boldsymbol{x}}_{\cdot t} - \bar{\boldsymbol{\xi}}_{\cdot t}) = \mathbf{0}_{K1}$, so that $\operatorname{plim}[\bar{\boldsymbol{x}}_{\cdot t}\bar{\varepsilon}_{\cdot t}] = \mathbf{0}_{K1}$ even if $\operatorname{plim}[\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_{it}\varepsilon_{it}] \neq \mathbf{0}_{K1}$. From (10.18) and (10.19) we therefore get

$$\operatorname{plim}[(\Delta_s \bar{\boldsymbol{x}}_{\cdot t})(\Delta_s \bar{y}_{\cdot t})] = \operatorname{plim}[(\Delta_s \bar{\boldsymbol{x}}_{\cdot t})(\Delta_s \bar{\boldsymbol{x}}'_{\cdot t})]\boldsymbol{\beta}, \tag{10.20}$$

$$\operatorname{plim}[(\bar{\boldsymbol{x}}_{\cdot t} - \bar{\boldsymbol{x}})(\bar{y}_{\cdot t} - \bar{y})] = \operatorname{plim}[(\bar{\boldsymbol{x}}_{\cdot t} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{\cdot t} - \bar{\boldsymbol{x}})']\boldsymbol{\beta}. \tag{10.21}$$

Hence, provided that $\mathrm{E}[(\Delta_s \bar{\boldsymbol{\xi}}_{\cdot t})(\Delta_s \bar{\boldsymbol{\xi}}_{\cdot t})']$ and $\mathrm{E}[(\bar{\boldsymbol{\xi}}_{\cdot t} - \bar{\boldsymbol{\xi}})(\bar{\boldsymbol{\xi}}_{\cdot t} - \bar{\boldsymbol{\xi}})']$ have rank K, which is ensured by Assumption (E), consistent estimators of $\boldsymbol{\beta}$ can be obtained by applying OLS on (10.18) and (10.19), which give, respectively,

$$\widehat{\boldsymbol{\beta}}_{\Delta s} = \left[\sum_{t=s+1}^{T} (\Delta_s \bar{\boldsymbol{x}}_{\cdot t})(\Delta_s \bar{\boldsymbol{x}}_{\cdot t})'\right]^{-1} \left[\sum_{t=s+1}^{T} (\Delta_s \bar{\boldsymbol{x}}_{\cdot t})(\Delta_s \bar{y}_{\cdot t})\right], \quad s = 1,\ldots,T-1, \tag{10.22}$$

$$\widehat{\boldsymbol{\beta}}_{\mathrm{BP}} = \left[\sum_{t=1}^{T} (\bar{\boldsymbol{x}}_{\cdot t} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{\cdot t} - \bar{\boldsymbol{x}})'\right]^{-1} \left[\sum_{t=1}^{T} (\bar{\boldsymbol{x}}_{\cdot t} - \bar{\boldsymbol{x}})(\bar{y}_{\cdot t} - \bar{y})\right]. \tag{10.23}$$

The latter is the 'between period' (BP) estimator. The consistency of these estimators simply relies on the fact that averages of a large number of repeated measurements of an error-ridden variable give, under weak conditions, an error-free measure of the true average at the limit, *provided that this average shows variation along the remaining dimension,* i.e.*, across periods*. Shalabh (2003) also discusses consistent coefficient estimation in measurement error models with replicated observations. The latter property is ensured by Assumption (E). A major problem with these estimators is their low potential efficiency, as none of them exploits the between individual variation in the data, which often is the main source of variation.

Basic to these conclusions is the assumption that *the measurement error has no period specific component*, which, roughly speaking, means that it is 'equally difficult' to measure $\boldsymbol{\xi}$ correctly in all periods. If such a component is present, it will not vanish when taking plims of period means, i.e., $\operatorname{plim}(\bar{\boldsymbol{v}}_{\cdot t})$ will no longer be zero, (10.20) and (10.21) will no longer hold, and so $\widehat{\boldsymbol{\beta}}_{\Delta s}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{BP}}$ will be inconsistent.

10.2.5 GMM Estimation and Testing in the General Case

We first consider the GMM principle in general, without reference to panel data and measurement error situations. Assume that we want to estimate the $(K \times 1)$

[3] Throughout plim denotes probability limits when N goes to infinity and T is finite.

coefficient vector $\boldsymbol{\beta}$ in the equation[4]

$$y = \boldsymbol{x}\boldsymbol{\beta} + \varepsilon, \tag{10.24}$$

where y and ε are scalars and $\boldsymbol{x}$ is a $(1 \times K)$ regressor vector. There exists an instrument vector $\boldsymbol{z}$, of dimension $(1 \times G)$, for $\boldsymbol{x}$ $(G \geq K)$, satisfying the OCs

$$\mathsf{E}(\boldsymbol{z}'\varepsilon) = \mathsf{E}[\boldsymbol{z}'(y - \boldsymbol{x}\boldsymbol{\beta})] = \mathbf{0}_{G1}. \tag{10.25}$$

We have n observations on $(y,\boldsymbol{x},\boldsymbol{z})$, denoted as $(y_j,\boldsymbol{x}_j,\boldsymbol{z}_j)$, $j = 1,\ldots,n$, and define the vector valued $(G \times 1)$ function of corresponding empirical means,

$$\boldsymbol{g}_n(y,\boldsymbol{x},z;\boldsymbol{\beta}) = \tfrac{1}{n}\textstyle\sum_{j=1}^{n} \boldsymbol{z}_j'(y_j - \boldsymbol{x}_j\boldsymbol{\beta}). \tag{10.26}$$

It may be considered the empirical counterpart to $\mathsf{E}[\boldsymbol{z}'(y - \boldsymbol{x}\boldsymbol{\beta})]$ based on the sample. The *essence of* GMM is to choose as an estimator for $\boldsymbol{\beta}$ the value which brings the value of $\mathbf{g}_n(y,\boldsymbol{x},z;\boldsymbol{\beta})$ as close to its theoretical counterpart, $\mathbf{0}_{G1}$, as possible. If $G = K$, an exact solution to $\mathbf{g}_n(y,\boldsymbol{x},z;\boldsymbol{\beta}) = \mathbf{0}_{G1}$ exists and is the simple IV estimator

$$\boldsymbol{\beta}^* = [\textstyle\sum_j \boldsymbol{z}_j'\boldsymbol{x}_j]^{-1}[\sum_j \boldsymbol{z}_j' y_j].$$

If $G > K$, which is the most common situation, GMM solves the estimation problem by *minimizing a distance measure represented by a quadratic form in* $g_n(y,\boldsymbol{x},z;\boldsymbol{\beta})$ *for a suitably chosen positive definit* $(G \times G)$ *weighting matrix* $\boldsymbol{W}_n$, i.e.,

$$\boldsymbol{\beta}^*_{\mathrm{GMM}} = \mathrm{argmin}_{\boldsymbol{\beta}}[\mathbf{g}_n(y,\boldsymbol{x},z;\boldsymbol{\beta})'\mathbf{W}_n\mathbf{g}_n(y,\boldsymbol{x},z;\boldsymbol{\beta})]. \tag{10.27}$$

All estimators obtained in this way are consistent. A choice which leads to an asymptotically efficient estimator of $\boldsymbol{\beta}$, is to set this weighting matrix equal (or proportional) to the inverse of (an estimate of) the (asymptotic) covariance matrix of $\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{z}_j'\varepsilon_j$; see, e.g., Davidson and MacKinnon (1993, Theorem 17.3) and Harris and Mátyás (1999, Sect. 1.3.3).

If ε is serially uncorrelated and homoskedastic, with variance σ_ε^2, the appropriate choice is simply $\mathbf{W}_n = [n^{-2}\sigma_\varepsilon^2 \sum_{j=1}^{n} \boldsymbol{z}_j'\boldsymbol{z}_j]^{-1}$. The estimator obtained from (10.27) is then

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\mathrm{GMM}} = {} & [(\textstyle\sum_j \boldsymbol{x}_j'\boldsymbol{z}_j)(\sum_j \boldsymbol{z}_j'\boldsymbol{z}_j)^{-1}(\sum_j \boldsymbol{z}_j'\boldsymbol{x}_j)]^{-1} \\ & \times [(\textstyle\sum_j \boldsymbol{x}_j'\boldsymbol{z}_j)(\sum_j \boldsymbol{z}_j'\boldsymbol{z}_j)^{-1}(\sum_j \boldsymbol{z}_j' y_j)],\end{aligned} \tag{10.28}$$

which is the standard Two-Stage Least Squares (2SLS) estimator. If ε_j has an unspecified heteroskedasticity or has a more or less strictly specified autocorrelation, we can reformulate the OCs in an appropriate way, as will be exemplified below. Both of these properties are essential for the application of GMM to panel data.

[4] We here, unlike in Sects. 10.2.1–10.2.4, let the column number denote the regressor and the row number the observation. Following this convention, we can express the following IV and GMM estimators in the more common format when going from vector to matrix notation.

To operationalize the latter method in the presence of unknown heteroskedasticity, we first construct consistent residuals $\widehat{\varepsilon}_j$, usually from (10.28), which we consider as a *first step GMM estimator*, and estimate $\mathbf{W}_n$ by $\widehat{\mathbf{W}}_n = [n^{-2}\sum_j \boldsymbol{z}_j'\widehat{\varepsilon}_j^{2}\boldsymbol{z}_j]^{-1}$; see White (1984, Sects. IV.3 and VI.2). Inserting this into (10.27) gives

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_{\mathrm{GMM}} = {} & [(\textstyle\sum_j \boldsymbol{x}_j'\boldsymbol{z}_j)(\sum_j \boldsymbol{z}_j'\widehat{\varepsilon}_j^2\boldsymbol{z}_j)^{-1}(\sum_j \boldsymbol{z}_j'\boldsymbol{x}_j)]^{-1} \\ & \times [(\textstyle\sum_j \boldsymbol{x}_j'\boldsymbol{z}_j)(\sum_j \boldsymbol{z}_j'\widehat{\varepsilon}_j^2\boldsymbol{z}_j)^{-1}(\sum_j \boldsymbol{z}_j' y_j)]. \end{aligned} \tag{10.29}$$

This *second step GMM estimator* is in a sense an optimal GMM estimator in the presence of unspecified error/disturbance heteroskedasticity.

The validity of the orthogonality condition (10.25) can be tested by the Sargan-Hansen statistic [confer Hansen (1982), Newey (1985), and Arellano and Bond (1991)], corresponding to the asymptotically efficient estimator $\widetilde{\boldsymbol{\beta}}_{\mathrm{GMM}}$:

$$J = [(\textstyle\sum_j \widehat{\varepsilon}_j'\boldsymbol{z}_j)(\sum_j \boldsymbol{z}_j'\widehat{\varepsilon}_j^2\boldsymbol{z}_j)^{-1}(\sum_j \boldsymbol{z}_j'\widehat{\varepsilon}_j)]^{-1}.$$

Under the null, J is asymptotically distributed as χ^2 with a number of degrees of freedom equal to the number of overidentifying restrictions, i.e., the number of orthogonality conditions less the number of coefficients estimated under the null.

The procedures for estimating standard errors of $\widehat{\boldsymbol{\beta}}_{\mathrm{GMM}}$ and $\widetilde{\boldsymbol{\beta}}_{\mathrm{GMM}}$ can be explained as follows. Express (10.24) and (10.25) as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \mathsf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \qquad \mathsf{E}(\boldsymbol{Z}'\boldsymbol{\varepsilon}) = \mathbf{0}, \qquad \mathsf{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\Omega},$$

where $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{\varepsilon}$ correspond to y, $\boldsymbol{x}$, z and ε, and the n observations are placed along the rows. The two generic GMM estimators (10.28) and (10.29) have the form

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= [\boldsymbol{X}'\boldsymbol{P}_Z\boldsymbol{X}]^{-1}[\boldsymbol{X}'\boldsymbol{P}_Z\boldsymbol{y}], & \boldsymbol{P}_Z &= \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}', \\ \widetilde{\boldsymbol{\beta}} &= [\boldsymbol{X}'\boldsymbol{P}_Z(\boldsymbol{\Omega})\boldsymbol{X}]^{-1}[\boldsymbol{X}'\boldsymbol{P}_Z(\boldsymbol{\Omega})\boldsymbol{y}], & \boldsymbol{P}_Z(\boldsymbol{\Omega}) &= \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{\Omega}\boldsymbol{Z})^{-1}\boldsymbol{Z}'. \end{aligned}$$

Let the residual vector obtained from the former be $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and

$$\begin{aligned} \boldsymbol{S}_{XZ} = \boldsymbol{S}_{ZX}' = \frac{\boldsymbol{X}'\boldsymbol{Z}}{n}, \quad \boldsymbol{S}_{ZZ} = \frac{\boldsymbol{Z}'\boldsymbol{Z}}{n}, \quad \boldsymbol{S}_{\varepsilon Z} = \boldsymbol{S}_{Z\varepsilon}' = \frac{\boldsymbol{\varepsilon}'\boldsymbol{Z}}{n}, \\ \boldsymbol{S}_{Z\boldsymbol{\Omega}Z} = \frac{\boldsymbol{Z}'\boldsymbol{\Omega}\boldsymbol{Z}}{n}, \quad \boldsymbol{S}_{Z\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}Z} = \frac{\boldsymbol{Z}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{Z}}{n}, \quad \boldsymbol{S}_{Z\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}Z} = \frac{\boldsymbol{Z}'\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}'\boldsymbol{Z}}{n}. \end{aligned}$$

Inserting for $\boldsymbol{y}$ in the expressions for the two estimators gives

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}[\boldsymbol{X}'\boldsymbol{P}_Z\boldsymbol{X}]^{-1}[\boldsymbol{X}'\boldsymbol{P}_Z\boldsymbol{\varepsilon}] = [\boldsymbol{S}_{XZ}\boldsymbol{S}_{ZZ}^{-1}\boldsymbol{S}_{ZX}]^{-1}\left[\boldsymbol{S}_{XZ}\boldsymbol{S}_{ZZ}^{-1}\frac{\boldsymbol{Z}'\boldsymbol{\varepsilon}}{\sqrt{n}}\right],$$

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}[\boldsymbol{X}'\boldsymbol{P}_Z(\boldsymbol{\Omega})\boldsymbol{X}]^{-1}[\boldsymbol{X}'\boldsymbol{P}_Z(\boldsymbol{\Omega})\boldsymbol{\varepsilon}] = [\boldsymbol{S}_{XZ}\boldsymbol{S}_{Z\boldsymbol{\Omega}Z}^{-1}\boldsymbol{S}_{ZX}]^{-1}\left[\boldsymbol{S}_{XZ}\boldsymbol{S}_{Z\boldsymbol{\Omega}Z}^{-1}\frac{\boldsymbol{Z}'\boldsymbol{\varepsilon}}{\sqrt{n}}\right],$$

and hence

$$n(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})' = [\boldsymbol{S}_{XZ}\boldsymbol{S}_{ZZ}^{-1}\boldsymbol{S}_{ZX}]^{-1}[\boldsymbol{S}_{XZ}\boldsymbol{S}_{ZZ}^{-1}\boldsymbol{S}_{Z\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}Z}\boldsymbol{S}_{ZZ}^{-1}\boldsymbol{S}_{ZX}][\boldsymbol{S}_{XZ}\boldsymbol{S}_{ZZ}^{-1}\boldsymbol{S}_{ZX}]^{-1},$$
$$n(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})' = [\boldsymbol{S}_{XZ}\boldsymbol{S}_{Z\Omega Z}^{-1}\boldsymbol{S}_{ZX}]^{-1}[\boldsymbol{S}_{XZ}\boldsymbol{S}_{Z\Omega Z}^{-1}\boldsymbol{S}_{Z\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}Z}\boldsymbol{S}_{Z\Omega Z}^{-1}\boldsymbol{S}_{ZX}][\boldsymbol{S}_{XZ}\boldsymbol{S}_{Z\Omega Z}^{-1}\boldsymbol{S}_{ZX}]^{-1}.$$

The asymptotic covariance matrices of $\sqrt{n}\widehat{\boldsymbol{\beta}}$ and $\sqrt{n}\widetilde{\boldsymbol{\beta}}$ can then, under suitable regularity conditions, be written as [see Bowden and Turkington (1984, pp. 26, 69)]

$$a\mathsf{V}(\sqrt{n}\widehat{\boldsymbol{\beta}}) = \lim \mathsf{E}[n(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'] = \operatorname{plim}[n(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'],$$
$$a\mathsf{V}(\sqrt{n}\widetilde{\boldsymbol{\beta}}) = \lim \mathsf{E}[n(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})'] = \operatorname{plim}[n(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta})'].$$

Since $\mathbf{S}_{Z\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}Z}$ and $\mathbf{S}_{Z\Omega Z}$ coincide asymptotically, we get, letting bars denote plims,

$$a\mathsf{V}(\sqrt{n}\widehat{\boldsymbol{\beta}}) = [\overline{\boldsymbol{S}}_{XZ}\overline{\boldsymbol{S}}_{ZZ}^{-1}\overline{\boldsymbol{S}}_{ZX}]^{-1}[\overline{\boldsymbol{S}}_{XZ}\overline{\boldsymbol{S}}_{ZZ}^{-1}\overline{\boldsymbol{S}}_{Z\Omega Z}\overline{\boldsymbol{S}}_{ZZ}^{-1}\overline{\boldsymbol{S}}_{ZX}][\overline{\boldsymbol{S}}_{XZ}\overline{\boldsymbol{S}}_{ZZ}^{-1}\overline{\boldsymbol{S}}_{ZX}]^{-1},$$
$$a\mathsf{V}(\sqrt{n}\widetilde{\boldsymbol{\beta}}) = [\overline{\boldsymbol{S}}_{XZ}\overline{\boldsymbol{S}}_{Z\Omega Z}^{-1}\overline{\boldsymbol{S}}_{ZX}]^{-1}.$$

Replacing the plims $\overline{\mathbf{S}}_{XZ}, \overline{\mathbf{S}}_{ZX}, \overline{\mathbf{S}}_{ZZ}$ and $\overline{\mathbf{S}}_{Z\Omega Z}$ by their sample counterparts, $\mathbf{S}_{XZ}, \mathbf{S}_{ZX}$, $\mathbf{S}_{ZZ}$ and $\mathbf{S}_{Z\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}Z}$ and dividing by n, we get the following estimators of the asymptotic covariance matrices of $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$:

$$\widehat{\mathsf{V}(\widehat{\boldsymbol{\beta}})} = \frac{1}{n}[\boldsymbol{S}_{XZ}\boldsymbol{S}_{ZZ}^{-1}\boldsymbol{S}_{ZX}]^{-1}[\boldsymbol{S}_{XZ}\boldsymbol{S}_{ZZ}^{-1}\boldsymbol{S}_{Z\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}Z}\boldsymbol{S}_{ZZ}^{-1}\boldsymbol{S}_{ZX}][\boldsymbol{S}_{XZ}\boldsymbol{S}_{ZZ}^{-1}\boldsymbol{S}_{ZX}]^{-1}$$
$$= [\boldsymbol{X}'\boldsymbol{P}_Z\boldsymbol{X}]^{-1}[\boldsymbol{X}'\boldsymbol{P}_Z\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}'\boldsymbol{P}_Z\boldsymbol{X}][\boldsymbol{X}'\boldsymbol{P}_Z\boldsymbol{X}]^{-1},$$
$$\widehat{\mathsf{V}(\widetilde{\boldsymbol{\beta}})} = \frac{1}{n}[\boldsymbol{S}_{XZ}\boldsymbol{S}_{Z\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}Z}^{-1}\boldsymbol{S}_{ZX}]^{-1} = [\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}]^{-1} = [\boldsymbol{X}'\boldsymbol{P}_Z(\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}')\boldsymbol{X}]^{-1}.$$

These are the generic expressions for estimating variances and covariances of the GMM estimators (10.28) and (10.29). When calculating $\widetilde{\boldsymbol{\beta}}$ in practice, we replace $\boldsymbol{P}_Z(\boldsymbol{\Omega})$ by $\boldsymbol{P}_Z(\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}') = \boldsymbol{Z}(\boldsymbol{Z}'\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ [see White (1982, 1984)].

10.2.6 Estimation by GMM, Combining Differences and Levels

Following this general description of the GMM, we can construct estimators of $\boldsymbol{\beta}$ by replacing the expectations in (10.14)–(10.17) by sample means taken over i and minimizing their distances from the zero vector. There are several ways in which this idea can be operationalized. We can

(i) Estimate equations in differences, with instruments in levels, using (10.14) and/or (10.15) for (a) one (t,θ) and one p, (b) one (t,θ) and several p, or (c) several (t,θ) and several p jointly.

(ii) Estimate equations in levels, with instruments in differences, using (10.16) and/or (10.17) for (a) one t and one (p,q), (b) one t and several (p,q), or (c) several t and several (p,q) jointly.

In cases (i.a) and (ii.a), we obtain an empirical distance equal to the zero vector, so no minimization is needed. This corresponds, formally, to the situation with 'exact identification' (exactly as many OCs as needed) in classical IV estimation. In cases (i.b), (i.c), (ii.b), and (ii.c), we have, in a formal sense, 'overidentification' (more than the necessary number of OCs), and therefore construct 'compromise estimators' by minimizing appropriate quadratic forms in the corresponding empirical distances.

We now consider cases (a), (b), and (c) for the differenced equation and the level equation.

(a) Simple period specific IV estimators

Equation in differences, IVs in levels. The sample mean counterpart to (10.14) and (10.15) for one (t,θ,p) gives the estimator

$$\widehat{\boldsymbol{\beta}}_{p(t\theta)} = [\textstyle\sum_{i=1}^{N} \boldsymbol{z}_{ip}(\Delta \boldsymbol{x}'_{it\theta})]^{-1}[\sum_{i=1}^{N} \boldsymbol{z}_{ip}(\Delta y_{it\theta})], \tag{10.30}$$

where $\boldsymbol{z}_{ip} = \boldsymbol{x}_{ip}$ or equal to $\boldsymbol{x}_{ip}$ with one element replaced by y_{ip}.

Equation in levels, IVs in differences. The sample mean counterpart to (10.16) and (10.17) for one (t,p,q) gives the estimator

$$\widehat{\boldsymbol{\beta}}_{(pq)t} = [\textstyle\sum_{i=1}^{N} (\Delta \boldsymbol{z}_{ipq})\boldsymbol{x}'_{it}]^{-1}[\sum_{i=1}^{N} (\Delta \boldsymbol{z}_{ipq}) y_{it}], \tag{10.31}$$

where $\Delta \boldsymbol{z}_{ipq} = \Delta \boldsymbol{x}_{ipq}$ or equal to $\Delta \boldsymbol{x}_{ipq}$ with one element replaced by Δy_{ipq}. Using (10.14)–(10.17) we note that

- When $\boldsymbol{z}_{ip} = \boldsymbol{x}_{ip}$ $(p \neq \theta, t)$ and $\Delta \boldsymbol{z}_{ipq} = \Delta \boldsymbol{x}_{ipq}$ $(t \neq p,q)$, Assumption (B2) is necessary for consistency of $\widehat{\boldsymbol{\beta}}_{p(t\theta)}$ and $\widehat{\boldsymbol{\beta}}_{(pq)t}$. If y_{ip} is included in $\boldsymbol{z}_{ip}$ $(p \neq \theta, t)$, and Δy_{pq} is included in $\Delta \boldsymbol{z}_{ipq}$ $(t \neq p,q)$, Assumption (C2) is also necessary for consistency of $\widehat{\boldsymbol{\beta}}_{p(t\theta)}$ and $\widehat{\boldsymbol{\beta}}_{(pq)t}$.
- Assumptions (D1) and (D2) are necessary for consistency of $\widehat{\boldsymbol{\beta}}_{(pq)t}$, but they are not necessary for consistency of $\widehat{\boldsymbol{\beta}}_{p(t\theta)}$.

Since the correlation between the regressors and the instruments, say between $\boldsymbol{z}_{ip}$ and $\Delta \boldsymbol{x}_{it\theta}$, may be low, (10.30) and (10.31) may suffer from the 'weak instrument problem', discussed in Nelson and Startz (1990), Davidson and MacKinnon (1993, pp. 217–224), and Staiger and Stock (1997). The following estimators may be an answer to this problem.

(b) Period specific GMM estimators

We next consider estimation of $\boldsymbol{\beta}$ in (10.4) for one pair of periods (t,θ), utilizing as IVs for $\Delta \boldsymbol{x}_{it\theta}$ *all admissible* $\boldsymbol{x}_{ip}$*s*, and estimation of $\boldsymbol{\beta}$ in (10.3), for one period (t), utilizing as IVs for $\boldsymbol{x}_{it}$ *all admissible* $\Delta \boldsymbol{x}_{ipq}$*s*.

To formalize this, we define the selection and differencing matrices

$$\boldsymbol{P}_{t\theta} = \begin{bmatrix} ((T-2)\times T) \text{ matrix} \\ \text{obtained by deleting from} \\ \text{the } T\text{-dimensional} \\ \text{identity matrix} \\ \text{rows } t \text{ and } \theta \end{bmatrix}, \quad \boldsymbol{D}_t = \begin{bmatrix} \boldsymbol{d}_{21} \\ \vdots \\ \boldsymbol{d}_{t-1,t-2} \\ \boldsymbol{d}_{t+1,t-1} \\ \boldsymbol{d}_{t+2,t+1} \\ \vdots \\ \boldsymbol{d}_{T,T-1} \end{bmatrix}, \quad t,\theta = 1,\ldots,T,$$

where $\boldsymbol{d}_{t\theta}$ is the $(1\times T)$ vector with element t equal to 1, element θ equal to -1 and zero otherwise, so that $\boldsymbol{D}_t$ is the are one-period $[(T-2)\times T)]$ differencing matrix, except that $\boldsymbol{d}_{t,t-1}$ and $\boldsymbol{d}_{t+1,t}$ are replaced by their sum, $\boldsymbol{d}_{t+1,t-1}$.[5] We use the notation

$$\begin{array}{ccc} \boldsymbol{y}_{i\cdot} = (y_{i1},\ldots,y_{iT})', & \boldsymbol{X}_{i\cdot} = (\boldsymbol{x}_{i1},\ldots,\boldsymbol{x}_{iT})', & \\ \boldsymbol{y}_{i(t\theta)} = \boldsymbol{P}_{t\theta}\boldsymbol{y}_{i\cdot}, & \boldsymbol{x}_{i(t\theta)} = \boldsymbol{P}_{t\theta}\boldsymbol{X}_{i\cdot}, & \boldsymbol{x}_{i(t\theta)} = \text{vec}(\boldsymbol{X}_{i(t\theta)})', \\ \Delta\boldsymbol{y}_{i(t)} = \boldsymbol{D}_t\boldsymbol{y}_{i\cdot}, & \Delta\boldsymbol{X}_{i(t)} = \boldsymbol{D}_t\boldsymbol{x}_{i\cdot}, & \Delta\boldsymbol{x}_{i(t)} = \text{vec}(\Delta\boldsymbol{X}_{i(t)})', \end{array}$$

etc. Here $\boldsymbol{X}_{i(t\theta)}$ denotes the $[(T-2)\times K]$ matrix of $\boldsymbol{x}$ levels obtained by *deleting* rows t and θ from $\boldsymbol{X}_{i\cdot}$, and $\Delta\boldsymbol{X}_{i(t)}$ denotes the $[(T-2)\times K]$ matrix of $\boldsymbol{x}$ differences obtained by stacking all one-period differences between rows of $\boldsymbol{X}_{i\cdot}$ *not including period* t and the single two-period difference between the columns for periods $t+1$ and $t-1$. The vectors $\boldsymbol{y}_{i(t\theta)}$ and $\Delta\boldsymbol{y}_{i(t)}$ are constructed from $\boldsymbol{y}_{i\cdot}$ in a similar way. Stacking $\boldsymbol{y}'_{i(t\theta)}$, $\Delta\boldsymbol{y}'_{i(t)}$, $\boldsymbol{x}_{i(t\theta)}$, and $\Delta\boldsymbol{x}_{i(t)}$, by individuals, we get

$$\boldsymbol{Y}_{(t\theta)} = \begin{bmatrix} \boldsymbol{y}'_{1(t\theta)} \\ \vdots \\ \boldsymbol{y}'_{N(t\theta)} \end{bmatrix}, \quad \Delta\boldsymbol{Y}_{(t)} = \begin{bmatrix} \Delta\boldsymbol{y}'_{1(t)} \\ \vdots \\ \Delta\boldsymbol{y}'_{N(t)} \end{bmatrix},$$

$$\boldsymbol{X}_{(t\theta)} = \begin{bmatrix} \boldsymbol{x}_{1(t\theta)} \\ \vdots \\ \boldsymbol{x}_{N(t\theta)} \end{bmatrix}, \quad \Delta\boldsymbol{X}_{(t)} = \begin{bmatrix} \Delta\boldsymbol{x}_{1(t)} \\ \vdots \\ \Delta\boldsymbol{x}_{N(t)} \end{bmatrix},$$

which have dimensions $(N\times(T-2))$, $(N\times(T-2))$, $(N\times(T-2)K)$, and $(N\times(T-2)K)$, respectively. These four matrices contain the IVs to be considered below.

Equation in differences, IVs in levels. Write (10.4) as

$$\Delta\boldsymbol{y}_{t\theta} = \Delta\boldsymbol{X}_{t\theta}\boldsymbol{\beta} + \Delta\boldsymbol{\varepsilon}_{t\theta},$$

where $\Delta\boldsymbol{y}_{t\theta} = (\Delta y_{1t\theta},\ldots,\Delta y_{Nt\theta})'$, $\Delta\boldsymbol{X}_{t\theta} = (\Delta\boldsymbol{x}_{1t\theta},\ldots,\Delta\boldsymbol{x}_{Nt\theta})'$, etc. Using $\boldsymbol{X}_{(t\theta)}$ as IV matrix for $\Delta\boldsymbol{X}_{t\theta}$, we obtain the following estimator of $\boldsymbol{\beta}$, *specific to period* (t,θ) *differences and utilizing all admissible* $\boldsymbol{x}$ *level IVs*,

[5] The two-period difference is effective only for $t = 2,\ldots,T-1$.

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{x(t\theta)} &= \left[(\Delta \boldsymbol{X}_{t\theta})'\boldsymbol{X}_{(t\theta)}\left(\boldsymbol{X}'_{(t\theta)}\boldsymbol{X}_{(t\theta)}\right)^{-1}\boldsymbol{X}'_{(t\theta)}(\Delta \boldsymbol{X}_{t\theta})\right]^{-1} \\ &\quad\times\left[(\Delta \boldsymbol{X}_{t\theta})'\boldsymbol{X}_{(t\theta)}\left(\boldsymbol{X}'_{(t\theta)}\boldsymbol{X}_{(t\theta)}\right)^{-1}\boldsymbol{X}'_{(t\theta)}(\Delta \boldsymbol{y}_{t\theta})\right] \\ &= \left[\left[\textstyle\sum_i(\Delta \boldsymbol{x}_{it\theta})\boldsymbol{x}'_{i(t\theta)}\right]\left[\textstyle\sum_i \boldsymbol{x}_{i(t\theta)}\boldsymbol{x}'_{i(t\theta)}\right]^{-1}\left[\textstyle\sum_i \boldsymbol{x}_{i(t\theta)}(\Delta \boldsymbol{x}'_{it\theta})\right]\right]^{-1} \\ &\quad\times\left[\left[\textstyle\sum_i(\Delta \boldsymbol{x}_{it\theta})\boldsymbol{x}'_{i(t\theta)}\right]\left[\textstyle\sum_i \boldsymbol{x}_{i(t\theta)}\boldsymbol{x}'_{i(t\theta)}\right]^{-1}\left[\textstyle\sum_i \boldsymbol{x}_{i(t\theta)}(\Delta y_{it\theta})\right]\right]. \qquad (10.32)\end{aligned}$$

It exists if $\boldsymbol{X}'_{(t\theta)}\boldsymbol{X}_{(t\theta)}$ has rank $(T-2)K$, which requires $N \geq (T-2)K$. This GMM estimator, which exemplifies (10.28), minimizes the quadratic form:

$$\left(\frac{1}{N}\boldsymbol{X}'_{(t\theta)}\Delta\boldsymbol{\varepsilon}_{t\theta}\right)'\left(\frac{1}{N^2}\boldsymbol{X}'_{(t\theta)}\boldsymbol{X}_{(t\theta)}\right)^{-1}\left(\frac{1}{N}\boldsymbol{X}'_{(t\theta)}\Delta\boldsymbol{\varepsilon}_{t\theta}\right).$$

The weight matrix $(N^{-2}\boldsymbol{X}'_{(t\theta)}\boldsymbol{X}_{(t\theta)})^{-1}$ is proportional to the inverse of the (asymptotic) covariance matrix of $N^{-1}\boldsymbol{X}'_{(t\theta)}\Delta\boldsymbol{\varepsilon}_{t\theta}$ when $\Delta\boldsymbol{\varepsilon}_{it\theta}$ is IID across i, possibly with a variance depending on (t,θ). The consistency of $\widehat{\boldsymbol{\beta}}_{x(t\theta)}$ relies on Assumptions (B2) and (E).

Interesting modifications of $\widehat{\boldsymbol{\beta}}_{x(t\theta)}$ are:

(1) If $\mathrm{var}(\Delta\varepsilon_{it\theta}) = \omega_{it\theta}$ varies with i and is known, we can increase the efficiency of (10.32) by replacing $\boldsymbol{x}'_{i(t\theta)}\boldsymbol{x}_{i(t\theta)}$ by $\boldsymbol{x}'_{i(t\theta)}\omega_{it\theta}\boldsymbol{x}_{i(t\theta)}$, which gives an asymptotically optimal GMM estimator.[6] $\sum_i \boldsymbol{x}'_{i(t\theta)}\omega_{it\theta}\boldsymbol{x}_{i(t\theta)}$ for unknown $\omega_{it\theta}$ proceeds as in (10.29).

(2) Instead of using $\boldsymbol{X}_{(t\theta)}$ as IV matrix for $\Delta\boldsymbol{X}_{t\theta}$, as in (10.32), we may use $(\boldsymbol{X}_{(t\theta)} \vdots \mathbf{Y}_{(t\theta)})$.

Equation in levels, IVs in differences. Write (10.3) as

$$\boldsymbol{y}_t = c\boldsymbol{e}_N + \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{t\theta},$$

where $\mathbf{e}_N$ is the N-vector of ones, $\boldsymbol{y}_t = (y_{1t},\ldots,y_{Nt})'$, $\boldsymbol{X}_t = (\boldsymbol{x}_{1t},\ldots,\boldsymbol{x}_{Nt})'$, etc. Using $\Delta\boldsymbol{X}_{(t)}$ as IV matrix for $\boldsymbol{X}_t$, we get the following estimator of $\boldsymbol{\beta}$, *specific to period t levels, utilizing all admissible* $\boldsymbol{x}$ *difference IVs*,

[6] For a more general treatment of asymptotic efficiency in estimation with moment conditions, see Chamberlain (1987) and Newey and McFadden (1994).

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{x(t)} &= \left[\boldsymbol{X}_t'(\Delta\boldsymbol{X}'_{(t)})\left((\Delta\boldsymbol{X}'_{(t)})(\Delta\boldsymbol{X}_{(t)})\right)^{-1}(\Delta\boldsymbol{X}_{(t)})'\boldsymbol{X}_t\right]^{-1} \\ &\quad\times\left[\boldsymbol{X}_t'(\Delta\boldsymbol{X}_{(t)})\left((\Delta\boldsymbol{X}_{(t)})'(\Delta\boldsymbol{X}_{(t)})\right)^{-1}(\Delta\boldsymbol{X}_{(t)})'\boldsymbol{y}_t\right] \\ &= \left[\left[\textstyle\sum_i \boldsymbol{x}_{it}(\Delta\boldsymbol{x}'_{i(t)})\right]\left[\textstyle\sum_i(\Delta\boldsymbol{x}_{i(t)})(\Delta\boldsymbol{x}'_{i(t)})\right]^{-1}\left[\textstyle\sum_i(\Delta\boldsymbol{x}_{i(t)})\boldsymbol{x}'_{it}\right]\right]^{-1} \\ &\quad\times\left[\left[\textstyle\sum_i \boldsymbol{x}_{it}(\Delta\boldsymbol{x}'_{i(t)})\right]\left[\textstyle\sum_i(\Delta\boldsymbol{x}_{i(t)})(\Delta\boldsymbol{x}'_{i(t)})\right]^{-1}\left[\textstyle\sum_i(\Delta\boldsymbol{x}_{i(t)})y_{it}\right]\right]. \quad (10.33)\end{aligned}$$

It exists if $(\Delta\boldsymbol{X}_{(t)})'(\Delta\boldsymbol{X}_{(t)})$ has rank $(T-2)K$, which again requires $N \geq (T-2)K$. This GMM estimator, which also exemplifies (10.28), minimizes the quadratic form:

$$\left(\frac{1}{N}(\Delta\boldsymbol{X}_{(t)})'\boldsymbol{\varepsilon}_t\right)'\left[\frac{1}{N^2}(\Delta\boldsymbol{X}_{(t)})'(\Delta\boldsymbol{X}_{(t)})\right]^{-1}\left(\frac{1}{N}(\Delta\boldsymbol{X}_{(t)})'\boldsymbol{\varepsilon}_t\right).$$

The weight matrix $[N^{-2}(\Delta\boldsymbol{X}_{(t)})'(\Delta\boldsymbol{X}_{(t)})]^{-1}$ is proportional to the inverse of the (asymptotic) covariance matrix of $N^{-1}(\Delta\boldsymbol{X}_{(t)})'\boldsymbol{\varepsilon}_t$ when ε_{it} is IID across i, possibly with a variance depending on t. The consistency of $\widehat{\boldsymbol{\beta}}_{x(t)}$ relies on (B3), (D1), (D2), and the validity of (E3) for all (p,q).

Interesting modifications of $\widehat{\boldsymbol{\beta}}_{x(t)}$ are:

(1) If $\mathrm{var}(\varepsilon_{it}) = \omega_{it}$ varies with i and is known, we can increase the efficiency of (10.33) by replacing $(\Delta\boldsymbol{x}_{i(t)})'(\Delta\boldsymbol{x}_{i(t)})$ by $(\Delta\boldsymbol{x}_{i(t)})'\omega_{it}(\Delta\boldsymbol{x}_{i(t)})$, which gives an asymptotically optimal GMM estimator. Estimation of $\sum_i(\Delta\boldsymbol{x}_{i(t)})'\omega_{it}(\Delta\boldsymbol{x}_{i(t)})$ for unknown ω_{it} proceeds as in (10.29).

(2) Instead of using $\Delta\boldsymbol{X}_{(t)}$ as IV matrix for $\boldsymbol{X}_t$, as in (10.33), we may use $(\Delta\boldsymbol{X}_{(t)} \vdots \Delta\mathbf{Y}_{(t)})$.

If we replace assumptions (B2) and (C2) by (B1) or (C1) with arbitrary τ, we must ensure that the IVs have a lead or lag of at least $\tau+1$ periods to the regressor, to 'get clear of' the τ period memory of the MA(τ) process. Formally, we then replace $\boldsymbol{P}_{t\theta}$ and $\boldsymbol{D}_t$ by[7]

$$\boldsymbol{P}_{t\theta(\tau)} = \begin{bmatrix}\text{matrix obtained by}\\ \text{deleting from}\\ \text{the } T\text{-dimensional}\\ \text{identity matrix}\\ \text{rows } \theta-\tau,\ldots,\theta+\tau\\ \text{and } t-\tau,\ldots,t+\tau\end{bmatrix}, \quad \boldsymbol{D}_{t(\tau)} = \begin{bmatrix}\mathbf{d}_{21}\\ \vdots\\ \mathbf{d}_{t-\tau-1,t-\tau-2}\\ \mathbf{d}_{t+\tau+1,t-\tau-1}\\ \mathbf{d}_{t+\tau+2,t+\tau+1}\\ \vdots\\ \mathbf{d}_{T,T-1}\end{bmatrix}, \quad t,\theta = 1,\ldots,T,$$

and otherwise proceed as above.

[7] The dimension of these matrices depends in general on τ.

(c) Composite GMM estimators

We finally consider GMM estimation of $\boldsymbol{\beta}$ when we combine all essential OCs delimited by Propositions 10.1 and 10.2. We here assume that either (B1) and (C1) with $\tau = 0$ or (B1) and (B2) are satisfied. If $\tau > 0$, we can proceed as above, but must ensure that the variables in the IV matrix have a lead or lag of at least $\tau+1$ periods to the regressor, to 'get clear of' the τ period memory of the MA(τ) process, confer Part **B** of Propositions 10.1 and 10.2.

Equation in differences, IVs in levels. Consider (10.5) for all $\theta = t-1$ and all $\theta = t-2$. These $(T-1)+(T-2)$ equations stacked for individual i read

$$\begin{bmatrix} \Delta y_{i21} \\ \Delta y_{i32} \\ \vdots \\ \Delta y_{i,T,T-1} \\ \Delta y_{i31} \\ \Delta y_{i42} \\ \vdots \\ \Delta y_{i,T,T-2} \end{bmatrix} = \begin{bmatrix} \Delta \boldsymbol{x}'_{i21} \\ \Delta \boldsymbol{x}'_{i32} \\ \vdots \\ \Delta \boldsymbol{x}'_{i,T,T-1} \\ \Delta \boldsymbol{x}'_{i31} \\ \Delta \boldsymbol{x}'_{i42} \\ \vdots \\ \Delta \boldsymbol{x}'_{i,T,T-2} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \Delta \varepsilon_{i21} \\ \Delta \varepsilon_{i32} \\ \vdots \\ \Delta \varepsilon_{i,T,T-1} \\ \Delta \varepsilon_{i31} \\ \Delta \varepsilon_{i42} \\ \vdots \\ \Delta \varepsilon_{i,T,T-2} \end{bmatrix}, \tag{10.34}$$

or, compactly,

$$\Delta \boldsymbol{y}_i = (\Delta \boldsymbol{X}_i)\boldsymbol{\beta} + \Delta \boldsymbol{\varepsilon}_i.$$

The IV matrix, according to Proposition 10.1, is the $((2T-3) \times KT(T-2))$ matrix[8]

$$\boldsymbol{Z}_i = \begin{bmatrix} \boldsymbol{x}_{i(21)} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{x}_{i(32)} & \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{x}_{i(T,T-1)} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{x}_{i2} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{x}_{i3} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{x}_{i,T-1} \end{bmatrix}. \tag{10.35}$$

Let

$$\Delta \boldsymbol{y} = [(\Delta \boldsymbol{y}_1)', \ldots, (\Delta \boldsymbol{y}_N)']', \quad \Delta \boldsymbol{\varepsilon} = [(\Delta \boldsymbol{\varepsilon}_1)', \ldots, (\Delta \boldsymbol{\varepsilon}_N)']',$$
$$\Delta \boldsymbol{X} = [(\Delta \boldsymbol{X}_1)', \ldots, (\Delta \boldsymbol{X}_N)']', \boldsymbol{Z} = [\boldsymbol{Z}'_1, \ldots, \boldsymbol{Z}'_N]'.$$

[8] Formally, we here use different IVs for the $(T-1)+(T-2)$ different equations in (10.4), with $\boldsymbol{\beta}$ as a common slope coefficient.

The GMM estimator corresponding to $\mathsf{E}[\boldsymbol{Z}_i'(\Delta\varepsilon_i)] = \mathbf{0}_{T(T-2)K,1}$, which minimizes $[N^{-1}(\Delta\boldsymbol{\varepsilon})'\boldsymbol{Z}](N^{-2}\boldsymbol{V})^{-1}[N^{-1}\boldsymbol{Z}'(\Delta\boldsymbol{\varepsilon})]$ for $\boldsymbol{V} = \boldsymbol{Z}'\boldsymbol{Z}$, can be written as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{Dx} &= \left[(\Delta\boldsymbol{X})'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\Delta\boldsymbol{X})\right]^{-1}\left[(\Delta\boldsymbol{X})'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\Delta\boldsymbol{y})\right] \\ &= \left[\left[\textstyle\sum_i(\Delta\boldsymbol{X}_i)'\boldsymbol{Z}_i\right]\left[\textstyle\sum_i\boldsymbol{Z}_i'\boldsymbol{Z}_i\right]^{-1}\left[\textstyle\sum_i\boldsymbol{Z}_i'(\Delta\boldsymbol{X}_i)\right]\right]^{-1} \\ &\quad\times\left[\left[\textstyle\sum_i(\Delta\boldsymbol{X}_i)'\boldsymbol{Z}_i\right]\left[\textstyle\sum_i\boldsymbol{Z}_i'\boldsymbol{Z}_i\right]^{-1}\left[\textstyle\sum_i\boldsymbol{Z}_i'(\Delta\boldsymbol{y}_i)\right]\right]. \qquad (10.36)\end{aligned}$$

It is possible to include not only the essential OCs, but also the redundant OCs when constructing this GMM estimator. The singularity of $\boldsymbol{Z}'\boldsymbol{Z}$ when including all OCs, due to the linear dependence between the redundant and the essential OCs, may be treated by replacing standard inverses in the estimation formulae by generalised (Moore-Penrose) inverses. The resulting estimator is $\widehat{\boldsymbol{\beta}}_{Dx}$, which is shown formally in Biørn and Klette (1998).

If $\Delta\boldsymbol{\varepsilon}$ has a non-scalar covariance matrix, a more efficient GMM estimator is obtained for $\boldsymbol{V} = \boldsymbol{V}_{Z(\Delta\varepsilon)} = \mathsf{E}[\boldsymbol{Z}'(\Delta\boldsymbol{\varepsilon})(\Delta\boldsymbol{\varepsilon})'\boldsymbol{Z}]$, which gives

$$\widetilde{\boldsymbol{\beta}}_{Dx} = \left[(\Delta\boldsymbol{X})'\boldsymbol{Z}\boldsymbol{V}_{Z(\Delta\varepsilon)}^{-1}\boldsymbol{Z}'(\Delta\boldsymbol{X})\right]^{-1}\left[(\Delta\boldsymbol{X})'\boldsymbol{Z}\boldsymbol{V}_{Z(\Delta\varepsilon)}^{-1}\boldsymbol{Z}'(\Delta\boldsymbol{y})\right]. \qquad (10.37)$$

We can estimate $\frac{1}{N}\boldsymbol{V}_{Z(\Delta\varepsilon)}$ consistently from the residuals obtained from (10.37), $\widehat{\Delta\boldsymbol{\varepsilon}}_i = \Delta\boldsymbol{y}_i - (\Delta\boldsymbol{X}_i)\widehat{\boldsymbol{\beta}}_{Dx}$, by means of [see White (1984, Sects. IV.3 and VI.2) and (1986, Sect. 3)]

$$\frac{\widehat{\boldsymbol{V}}_{Z(\Delta\varepsilon)}}{N} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Z}_i'(\widehat{\Delta\boldsymbol{\varepsilon}}_i)(\widehat{\Delta\boldsymbol{\varepsilon}}_i)'\boldsymbol{Z}_i. \qquad (10.38)$$

Inserting (10.38) in (10.37), we get the asymptotically optimal (feasible) GMM estimator[9]

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_{Dx} &= \left[\left[\textstyle\sum_i(\Delta\boldsymbol{X}_i)'\boldsymbol{Z}_i\right]\left[\textstyle\sum_i\boldsymbol{Z}_i'\widehat{\Delta\boldsymbol{\varepsilon}}_i\widehat{\Delta\boldsymbol{\varepsilon}}_i'\boldsymbol{Z}_i\right]^{-1}\left[\textstyle\sum_i\boldsymbol{Z}_i'(\Delta\boldsymbol{X}_i)\right]\right]^{-1} \\ &\quad\times\left[\left[\textstyle\sum_i(\Delta\boldsymbol{X}_i)'\boldsymbol{Z}_i\right]\left[\textstyle\sum_i\boldsymbol{Z}_i'\widehat{\Delta\boldsymbol{\varepsilon}}_i\widehat{\Delta\boldsymbol{\varepsilon}}_i'\boldsymbol{Z}_i\right]^{-1}\left[\textstyle\sum_i\boldsymbol{Z}_i'(\Delta\boldsymbol{y}_i)\right]\right]. \qquad (10.39)\end{aligned}$$

These estimators can be modified by extending in (10.37) all $\boldsymbol{x}_{i(t,t-1)}$ to $(\boldsymbol{x}_{i(t,t-1)} \vdots \boldsymbol{y}'_{i(t,t-1)})$ and all $\boldsymbol{x}_{it}$ to $(\boldsymbol{x}_{it} \vdots \boldsymbol{y}_{it})$, which also exploit the OCs in the ys.

Equation in levels, IVs in differences. Consider next the T stacked level equations for individual i [confer (10.3)]

$$\begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT}\end{bmatrix} = \begin{bmatrix} c \\ \vdots \\ c\end{bmatrix} + \begin{bmatrix} \boldsymbol{x}_{i1} \\ \vdots \\ \boldsymbol{x}_{iT}\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT}\end{bmatrix}, \qquad (10.40)$$

[9] It is possible to include the redundant OCs also when constructing this GMM estimator. Using generalised (Moore-Penrose) inverses, the estimator remains the same.

or, compactly,

$$\boldsymbol{y}_i = \mathbf{e}_T c + \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i.$$

The IV matrix, according to Proposition 10.2, is the $(T \times T(T-2)K)$ matrix[10]

$$\Delta\boldsymbol{Z}_i = \begin{bmatrix} \Delta\boldsymbol{x}_{i(1)} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \Delta\boldsymbol{x}_{i(T)} \end{bmatrix}. \tag{10.41}$$

Let

$$\begin{aligned} &\boldsymbol{y} = [\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N']', \quad \boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_N']', \\ &\boldsymbol{X} = [\boldsymbol{X}_1', \ldots, \boldsymbol{X}_N']', \; \Delta\boldsymbol{Z} = [(\Delta\boldsymbol{Z}_1)', \ldots, (\Delta\boldsymbol{Z}_N)']'. \end{aligned}$$

The GMM estimator corresponding to $\mathsf{E}[(\Delta\boldsymbol{Z}_i)'\boldsymbol{\varepsilon}_i] = \mathbf{0}_{T(T-2)K,1}$, which minimizes

$$[N^{-1}\boldsymbol{\varepsilon}'(\Delta\boldsymbol{Z})](N^{-2}\boldsymbol{V}_\Delta)^{-1}[N^{-1}(\Delta\boldsymbol{Z})'\boldsymbol{\varepsilon}] \quad \text{for} \quad \boldsymbol{V}_\Delta = (\Delta\boldsymbol{Z})'(\Delta\boldsymbol{Z}),$$

can be written as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{Lx} &= \left[\boldsymbol{X}'(\Delta\boldsymbol{Z})[(\Delta\boldsymbol{Z})'(\Delta\boldsymbol{Z})]^{-1}(\Delta\boldsymbol{Z})'\boldsymbol{X}\right]^{-1} \\ &\quad \times \left[\boldsymbol{X}'(\Delta\boldsymbol{Z})[(\Delta\boldsymbol{Z})'(\Delta\boldsymbol{Z})]^{-1}(\Delta\boldsymbol{Z})'\boldsymbol{y}\right] \\ &= \left[[\textstyle\sum_i \boldsymbol{X}_i'(\Delta\boldsymbol{Z}_i)]\,[\sum_i(\Delta\boldsymbol{Z}_i)'(\Delta\boldsymbol{Z}_i)]^{-1}\,[\sum_i(\Delta\boldsymbol{Z}_i)'\boldsymbol{X}_i]\right]^{-1} \\ &\quad \times \left[[\textstyle\sum_i \boldsymbol{X}_i'(\Delta\boldsymbol{Z}_i)]\,[\sum_i(\Delta\boldsymbol{Z}_i)'(\Delta\boldsymbol{Z}_i)]^{-1}\,[\sum_i(\Delta\boldsymbol{Z}_i)'\boldsymbol{y}_i]\right]. \end{aligned} \tag{10.42}$$

If $\boldsymbol{\varepsilon}$ has a non-scalar covariance matrix, a more efficient GMM estimator is obtained for $\boldsymbol{V}_\Delta = \boldsymbol{V}_{(\Delta Z)\varepsilon} = \mathsf{E}[(\Delta\boldsymbol{Z})'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'(\Delta\boldsymbol{Z})]$, which gives

$$\widetilde{\boldsymbol{\beta}}_{Lx} = \left[\boldsymbol{X}'(\Delta\boldsymbol{Z})\boldsymbol{V}^{-1}_{(\Delta Z)\varepsilon}(\Delta\boldsymbol{Z})'\boldsymbol{X}\right]^{-1}\left[\boldsymbol{X}'(\Delta\boldsymbol{Z})\boldsymbol{V}^{-1}_{(\Delta Z)\varepsilon}(\Delta\boldsymbol{Z})'\boldsymbol{y}\right]. \tag{10.43}$$

We can estimate $\frac{1}{N}\boldsymbol{V}_{(\Delta Z)\varepsilon}$ consistently from the residuals obtained from (10.43), by

$$\frac{\widehat{\boldsymbol{V}}_{(\Delta Z)\varepsilon}}{N} = \frac{1}{N}\sum_{i=1}^{N}(\Delta\boldsymbol{Z}_i)'\widehat{\boldsymbol{\varepsilon}}_i\widehat{\boldsymbol{\varepsilon}}_i'(\Delta\boldsymbol{Z}_i). \tag{10.44}$$

Inserting (10.44) in (10.44), we get the asymptotically optimal (feasible) GMM estimator

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}_{Lx} &= \left[[\textstyle\sum_i \boldsymbol{X}_i'(\Delta\boldsymbol{Z}_i)]\left[\sum_i(\Delta\boldsymbol{Z}_i)'\widehat{\boldsymbol{\varepsilon}}_i\widehat{\boldsymbol{\varepsilon}}_i'(\Delta\boldsymbol{Z}_i)\right]^{-1}[\sum_i(\Delta\boldsymbol{Z}_i)'\boldsymbol{X}_i]\right]^{-1} \\ &\quad \times \left[[\textstyle\sum_i \boldsymbol{X}_i'(\Delta\boldsymbol{Z}_i)]\left[\sum_i(\Delta\boldsymbol{Z}_i)'\widehat{\boldsymbol{\varepsilon}}_i\widehat{\boldsymbol{\varepsilon}}_i'(\Delta\boldsymbol{Z}_i)\right]^{-1}[\sum_i(\Delta\boldsymbol{Z}_i)'\boldsymbol{y}_i]\right]. \end{aligned} \tag{10.45}$$

[10] Again, we formally use different IVs for different equations, considering (10.40) as T different equations with $\boldsymbol{\beta}$ as a common slope coefficient.

These estimators can be modified by extending all $\Delta \boldsymbol{x}_{i(t)}$ to $(\Delta \boldsymbol{x}_{i(t)} \vdots \Delta \boldsymbol{y}'_{i(t)})$ in (10.41), which also exploit the OCs in the Δys. Other moment estimators, which will not be discussed specifically in the present EIV context, are considered for situations with predetermined IVs in Ziliak (1997), with the purpose of reducing the finite sample bias of asymptotically optimal GMM estimators.

10.2.7 Extensions: Modifications

All the methods presented so far rely on differencing as a way of eliminating the individual effects, either in the equation or in the instruments. This is convenient for the case where the individual heterogeneity has an unspecified correlation with the latent regressor vector and for the fixed effects case. Other ways of eliminating this effect in such situations are discussed in Wansbeek (2001). Their essence is to stack the matrix of covariances between the regressand and the regressors and eliminating these nuisance parameters by suitable projections. Exploiting a possible structure, suggested by our theory, on the covariance matrix of the $\boldsymbol{\xi}_{it}$s and α_i across individuals and periods, may lead to further extensions. Additional exploitable structure may be found in the covariance matrix of the y_{it}s. The latter will, however, lead to moment restrictions that are quadratic in the coefficient vector $\boldsymbol{\beta}$. Under non-normality, higher order moments may also, in principle, be exploited to improve efficiency, but again at the cost of a mathematically less tractable problem.

In a *random effects situation*, with zero correlation between $\boldsymbol{\xi}_{it}$ and α_i, and hence between $\boldsymbol{x}_{it}$ and α_i, differencing or projecting out the α_is will not be efficient, since they will not exploit this zero correlation. The GLS estimator, which would have been the minimum variance linear unbiased estimator in the absence of measurement errors, will no longer, in general, be consistent [see Biørn (1996, Sect. 10.4.3)], so it has to be modified. Finally, if the equation contains strongly exogenous regressors in addition to the error-contaminated ones, further moment conditions exist, which can lead to improved small sample efficiency of the GMM estimators. An improvement of small sample efficiency may also be obtained by replacing IV or GMM by LIML estimation; see Wansbeek and Meijer (2000, Sect. 6.6).

10.2.8 Concluding Remarks

Above we have demonstrated that several, rather simple, GMM estimators which may handle jointly the heterogeneity problem and the measurement error problem in panel data, exist. These problems may be ‘intractable’ when only pure (single or repeated) cross section data or pure time series data are available. Estimators using either equations in differences with level values as instruments, or equations in levels with differenced values as instruments are useful. In both cases, the differences may be taken over one period or more.

Even for the static model considered here, instruments constructed from the regressors (xs) as well as from the regressands (ys) may be of interest. GMM estimators combining both instrument sets in an optimal way are usually more precise than those using either of them. Although a substantial number of orthogonality conditions constructed from differences taken over two periods or more are redundant, adding the essential two-period difference orthogonality conditions to the one-period conditions in the GMM algorithm may significantly affect the result [confer the examples in Biørn (2000)].

Using levels as instruments for differences, or vice versa, as a general estimation strategy within a GMM framework, however, may raise problems related to 'weak instruments'. Finding operational ways of identifying such instruments among those utilizing essential orthogonality conditions in order to reduce their potential damage with respect to inefficiency, is a challenge for future research.

10.3 Simultaneity and Panel Data

Simultaneous equation models (SEM) or structural models as they are also sometimes called, have been around in the economic literature for a long time dating back to the period when the Econometric Society itself was formed. In spite of this long history, their relevance in modelling economic phenomena has not diminished; if at all it is only growing over time with the realisation that there is a high degree of interdependence among the different variables involved in the explanation of any socio-economic phenomenon.

The theory of simultaneous equations has become a must in any econometric course whatever level it may be. This is due to the fact any researcher needs to be made attentive to the potential endogenous regressor problem, be it in a single equation model or in a system of equations and this is the problem that the SEM theory precisely deals with.

At this stage it may be useful to distinguish between interdependent systems i.e. simultaneous equations and what are called systems of regression equations or seemingly unrelated regressions (SUR) in which there are no endogenous variables on the right hand side but non-zero correlations are assumed between error terms of different equations. We will see later in the section that the reduced form of a SEM is a special case of SUR.

In a panel data setting, in addition to the simultaneous nature of the model which invariably leads to non-zero correlation between the right hand side variables and the residual disturbance term, there is also the possibility of the same variables being correlated with the specific effects. However unlike in the correlated regressors case of Chap. 4 eliminating the specific effect alone does not solve the problem here and we need a more comprehensive approach to tackle it. We will develop generalizations of the two stage least squares (2SLS) and three stage least squares (3SLS) methods that are available in the classical SEM case. These generalizations can also be presented in a GMM framework, giving the corresponding optimal estimation in this context.

The most commonly encountered panel data SEM is the SEM with error component (EC) structure. Thus a major part of this chapter will the devoted to this extension and all its variants. Other generalizations will be briefly discussed at the end.

10.3.1 SEM with EC

10.3.1.1 The Model

This model proposes to account for the temporal and cross-sectional heterogeneity of panel data by means of an error components structure in the structural equations of a simultaneous equation system. In other words, the specific effects associated with pooled data are incorporated in an additive manner in the random element of each equation.

Let us consider a complete linear system of M equations in M current endogenous variables and K exogenous variables. We do not consider the presence of lagged endogenous variables in the system. The reader is referred to the separate chapter of this book dealing with dynamic panel data models for treatment of such cases.

By a 'complete' system, we assume that there are as many equations as there are endogenous variables and hence the system can be solved to obtain the reduced form. Further, we also assume that the data set is balanced i.e. observations are available for all the variables for all the units at all dates. Once again, the case of unbalanced panel data sets is dealt with in a separate chapter of the book.

We write the M–th structural equation of the system as follows:[11]

$$y'_{it}\gamma^*_m + x'_{it}\beta^*_m + u_{mit} = 0, \quad m = 1,\ldots M \tag{10.46}$$

where y'_{it} is the $(1 \times M)$ vector of observations on all the M endogenous variables for the i-th individual at the t-th time period; x'_{it} is the $(1 \times K)$ vector of observations on all the K exogenous variables for the i-th individual at the t-th time period; γ^*_m and β^*_m are respectively the coefficient vectors of y'_{it} and x'_{it}; and u_{mit} is the disturbance term of the m-th equation for the i-th individual and the t-th time period.

More explicitly,

$$y'_{it} = [y_{1it}\ldots y_{Mit}]; \quad x'_{it} = [x_{1it}\ldots x_{Kit}];$$

$$\beta^{*'}_m = [\beta^*_{1m}\ldots\beta^*_{Mm}]; \quad \gamma^{*'}_m = [\gamma^*_{1m}\ldots\gamma^*_{Km}]\ .$$

[11] Note that the constant term is included in the β vector, contrary to the introductory chapters, and hence x_{it} contains 1 as its first element.

By piling up all the observations in the following way:

$$Y = \begin{pmatrix} y'_{11} \\ \vdots \\ y'_{1T} \\ \vdots \\ y'_{NT} \end{pmatrix} \quad ; \quad X = \begin{pmatrix} x'_{11} \\ \vdots \\ x'_{1T} \\ \vdots \\ x'_{NT} \end{pmatrix} \quad ; \quad u_m = \begin{pmatrix} u_{m11} \\ \vdots \\ u_{m1T} \\ \vdots \\ u_{mNT} \end{pmatrix} ,$$

equation (10.46) can be written as:

$$Y\gamma_m^* + X\beta_m^* + u_m = 0, \quad m = 1, \ldots M \tag{10.47}$$

Defining

$$\Gamma = [\gamma_1^* \ldots \gamma_M^*]; \quad B = [\beta_1^* \ldots \beta_M^*]; \quad U = [u_1 \ldots u_M] ,$$

we can write the whole system of M equations as:

$$Y\Gamma + XB + U = 0 . \tag{10.48}$$

Before turning to the error structure, we add that the elements of Γ and B satisfy certain *a priori* restrictions, crucial for identification, in particular the normalisation rule ($\gamma_{ii}^* = -1$) and the exclusion restrictions (some elements of Γ and B are identically zero).

Following an error components pattern, it is assumed that each structural equation error u_{mit} is composed of three components: an individual effect μ_{mi}, a time effect ε_{mt} and a residual error ν_{mit}. Formally, we have:

Assumption 1:

$$\begin{aligned} u_{mit} = \mu_{mi} + \varepsilon_{mt} + \nu_{mit}, \qquad & m = 1, \ldots, M \\ & i = 1, \ldots, N \\ & t = 1, \ldots, T . \end{aligned} \tag{10.49}$$

By denoting

$$l'_T(1 \times T) = [1 \ \ldots \ 1]; \ \ \mu_m = \begin{pmatrix} \mu_{m1} \\ \vdots \\ \mu_{mN} \end{pmatrix} \ ; \ \ \varepsilon_m = \begin{pmatrix} \varepsilon_{m1} \\ \vdots \\ \varepsilon_{mT} \end{pmatrix} \ ; \ \ \nu_m = \begin{pmatrix} \nu_{m11} \\ \vdots \\ \nu_{m1T} \\ \vdots \\ \nu_{mNT} \end{pmatrix} ,$$

the above decomposition (10.49) can be written for all the observations, as:

$$u_m = (I_N \otimes l_T)\mu_m + (l_N \otimes I_T)\varepsilon_m + \nu_m, \quad m = 1, \ldots, M .$$

Assumption 2:

$$E(\mu_m) = 0; \quad E(\varepsilon_m) = 0; \quad E(\nu_m) = 0\,, \quad m = 1,\ldots,M\,.$$

Assumption 3:

$$\begin{aligned} E(\mu_m \mu'_{m'}) &= \sigma_{\mu mm'} I_N\,, \qquad m,m' = 1,\ldots,M \\ E(\varepsilon_m \varepsilon'_{m'}) &= \sigma_{\varepsilon mm'} I_T, \qquad m,m' = 1,\ldots,M \\ E(\nu_m \nu'_{m'}) &= \sigma_{\nu mm'} I_{NT}, \qquad m,m' = 1,\ldots,M \end{aligned}$$

Assumption 4:

$$E(\mu_m \varepsilon'_{m'}) = 0; \;\; E(\mu_m \nu'_{m'}) = 0; \;\; E(\varepsilon_m \nu'_{m'}) = 0, \;\; \forall m,m'\,.$$

We will also assume *independence*, two by two, among the different components whenever required, and normality of their distribution for ML estimation.

Assumption 5:

The error components are independent of the exogenous variables.

From these assumptions, the covariance matrix between u_m and $u_{m'}$, denoted as $\Sigma_{mm'}$, can be derived as:

$$\Sigma_{mm'} = E(u_m u'_{m'}) = \sigma_{\mu mm'}(I_N \otimes l_T l'_T) + \sigma_{\varepsilon mm'}(l_N l'_N \otimes I_T) + \sigma_{\nu mm'} I_{NT}. \quad (10.50)$$

The spectral decomposition of $\Sigma_{mm'}$ is given by (see Nerlove (1971))

$$\Sigma_{mm'} = \sigma_{1mm'} M_1 + \sigma_{2mm'} M_2 + \sigma_{3mm'} M_3 + \sigma_{4mm'} M_4 \quad (10.51)$$

where

$$\begin{aligned} \sigma_{1mm'} &= \sigma_{\nu mm'} \\ \sigma_{2mm'} &= \sigma_{\nu mm'} + T\sigma_{\mu mm'} \\ \sigma_{3mm'} &= \sigma_{\nu mm'} + N\sigma_{\varepsilon mm'} \\ \sigma_{4mm'} &= \sigma_{\nu mm'} + T\sigma_{\mu mm'} + N\sigma_{\varepsilon mm'} \end{aligned} \quad (10.52)$$

and

$$\begin{aligned} M_1 &= I_{NT} - \frac{1}{T}(I_N \otimes l_T l'_T) - \frac{1}{N}(l_N l'_N \otimes I_T) + \frac{1}{NT} l_{NT} l'_{NT} \\ &\quad \text{of rank} \quad m_1 = (N-1)(T-1); \\ M_2 &= \frac{1}{T}(I_N \otimes l_T l'_T) - \frac{1}{NT} l_{NT} l'_{NT} \\ &\quad \text{of rank} \quad m_2 = N-1; \\ M_3 &= \frac{1}{N}(l_N l'_N \otimes I_T) - \frac{1}{NT} l_{NT} l'_{NT} \\ &\quad \text{of rank} \quad m_3 = T-1; \\ M_4 &= \frac{1}{NT} l_{NT} l'_{NT} \\ &\quad \text{of rank} \quad m_4 = 1 \end{aligned}$$

with

$$\sum_{i=1}^{4} M_i = I_{NT} \; ; \quad M_i M_j = \delta_{ij} M_i \; .$$

Further, we note that

$$l'_{NT} M_i = 0 \; , \quad i = 1, \, 2, \, 3 \; ; \quad l'_{NT} M_4 = l'_{NT} .$$

By denoting

$$\Sigma_\mu = [\sigma_{\mu mm'}] \; , \Sigma_\varepsilon = [\sigma_{\varepsilon mm'}] \; , \Sigma_\nu = [\sigma_{\nu mm'}] \; , m, m' = 1, \ldots, M,$$

relations (10.52) can be written in matrix form as:

$$\Sigma_1 = \Sigma_\nu \; ; \Sigma_2 = \Sigma_\nu + T\Sigma_\mu \; ; \Sigma_3 = \Sigma_\nu + N\Sigma_\varepsilon \; ; \Sigma_4 = \Sigma_\nu + T\Sigma_\mu + N\Sigma_\varepsilon \; .$$

Note that $\Sigma_\mu, \Sigma_\varepsilon$ and Σ_ν are uniquely determined from $\Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4$ and vice versa.

Finally, the variance–covariance matrix of the structural form can be verified to be:

$$\Sigma = E((\text{vec } U)(\text{vec } U)') = \sum_{i=1}^{4} \Sigma_i \otimes M_i \tag{10.53}$$

with

$$\Sigma_i = [\sigma_{imm'}] \;\; m, m' = 1, \ldots M \quad \text{for} \quad i = 1, \, 2, \, 3, \, 4.$$

The inverse and determinant of Σ (useful for the estimation procedures of later sections) are given by (see Baltagi (1981) or Balestra and Krishnakumar (1987)):

$$\Sigma^{-1} = \sum_{i=1}^{4} \Sigma_i^{-1} \otimes M_i \; ; \;\; | \, \Sigma \, | = \; \Pi_{i=1}^{4} \, | \, \Sigma_i \, |^{m_i} \;\; . \tag{10.54}$$

10.3.1.2 The Reduced Form and the Identification Problem

By definition, the reduced form of a system of simultaneous equations is the solution of the system for the endogenous variables in terms of the exogenous variables and the disturbances. For our model, it is given by:

$$Y = X\Pi + V$$

where

$$\Pi = -B\Gamma^{-1} \; ; \;\; V = -U\Gamma^{-1} \; .$$

By using the properties of vec, we can write

$$\text{vec } V = (-\Gamma^{-1'} \otimes I) \;\; \text{vec } U$$

and thus we have:

$$E(\text{vec } V) = 0$$

and

$$\begin{aligned}\Omega &= E((\text{vec } V)(\text{vec } V)') \\ &= (-\Gamma^{-1'} \otimes I)\Sigma(-\Gamma^{-1} \otimes I) \\ &= \sum_{i=1}^{4} \Gamma^{-1'}\Sigma_i\Gamma^{-1} \otimes M_i \\ &= \sum_{i=1}^{4} \Omega_i \otimes M_i\end{aligned}$$

where

$$\Omega_i = \Gamma^{-1'}\Sigma_i\Gamma^{-1}\,, \quad i = 1,\ 2,\ 3,\ 4.$$

It can be easily verified that each reduced form equation has a three components error structure like any structural equation and the covariances across different reduced form equations are also of the same nature as those across different structural equations. However, an important point in which the reduced form differs from the structural form is that the right hand side variables of the former are uncorrelated with the errors whereas it is not the case in the latter due to simultaneity.

Thus the reduced form is a seemingly unrelated regression (SUR) model with error components. This model was originally proposed by Avery (1977) and is an important extension of panel data specifications to systems of equations. Our reduced form is in fact a special case of such a model as the explanatory variables are the same in each equation. Avery (1977) treated a more general case in which each equation has its own set of explanatory variables. This interpretation of our reduced form enables us to provide an interesting application of Avery's model combining SUR with error components (EC). We do not intend to go into the details of the inference procedures for the reduced form for want of space. In general, both ML and feasible GLS can be applied. Both are consistent, asymptotically normal and equivalent. The reader is referred to Krishnakumar (1988) for detailed derivations.

In the context of any simultaneous equation model, it is important to consider the problem of identification prior to estimation. In the case of the classical simultaneous equation model (with homoscedastic and non–auto–correlated errors), there is abundant literature on identification (see, for instance, Koopmans (1953), Fisher (1966), Rothenberg (1971), and Hausman and Taylor (1983)).

In our case of SEM with EC, as long as there are no a *priori* restrictions on the structural variances and covariances (i.e. no 'covariance restrictions' in the terminology of Hausman and Taylor), the identification problem is exactly the same as that of the classical model. In other words, in such a situation, we can separate the discussion on the identification of Γ and B from that of the Σ_i (s), $i = \mu,\ \varepsilon,\ \nu$. Thus, we would have the same rank and order conditions of identifiability of the elements of Γ and B, and the same definitions of under-identified, just-identified and over-identified equations. Once the structural coefficients are identified, the identification of the structural variance–covariance matrices is immediate, through the equations relating them to the reduced form covariance matrices.

Now, if we impose additional a priori restrictions on the structural variances and covariances, then it is no longer possible to separate the equations relating (Γ, B) to Π from those relating Σ_i(s) to Ω_i(s), $i = \mu,\ \varepsilon,\ \nu$ and one has to study the existence and uniqueness of solutions for the full system consisting of all the identifying equations, given the prior restrictions. This has been done for the classical simultaneous equation model by Hausman and Taylor (1983). One can follow the same approach for our model but one has to keep in mind the fact that, in the classical case, there is only one Σ whereas in our case there are three of these sets of relations: $\Omega_i \Gamma'^{-1} \Sigma_i$, $i = \mu,\ \varepsilon,\ \nu$.

One type of a priori covariance restrictions that do not need any particular analysis is that either Σ_μ or Σ_ε is identically equal to zero (i.e. only one specific effect is present in the model) and hence is identified. Note that, in this case, the corresponding Ω matrix (Ω_μ or Ω_ε) is also zero and the spectral decomposition of Σ (and Ω) is reduced to two terms only.

10.3.1.3 Structural Form Estimation

Generalised Two Stage Least Squares

Let us consider a structural equation, say the m–th one and write it as:

$$y_m = Y_m \gamma_m + X_m \beta_m + u_m \ , \tag{10.55}$$

in which the normalisation rule ($\beta^*_{mm} = -1$) and the exclusion restrictions are already substituted. We assume that these are the only a priori information available. Note that Y_m and X_m denote the matrices of observations on the M^*_m included endogenous and K_m included exogenous right hand side variables respectively and γ_m and β_m denote their respective coefficients. By defining

$$Z_m = [Y_m\ X_m]; \quad \alpha_m = \begin{pmatrix} \gamma_m \\ \beta_m \end{pmatrix},$$

we can rewrite (10.55) as

$$y_m = Z_m\ \alpha_m + u_m \tag{10.56}$$

and we recall (see (10.51)) that

$$E(u_m\ u'_{m'}) = \Sigma_{mm'} = \sum_{i=1}^{4} \sigma_{imm'} M_i \ .$$

The endogenous right hand side variables of (10.55) are correlated with both the individual effects and the residual error term. Hence classical methods like the OLS, GLS, or within will all yield inconsistent estimators and an appropriate procedure

is given by the IV method which typically consists in premultiplying the equation in question by a matrix of valid instruments and then applying GLS to the transformed equation.

In the classical case, the instrument for Z_m is taken to be X (see, for instance, Theil (1971)). In our case, it can be shown that, of all the transformations of X, say FX, the one which minimises the asymptotic variance–covariance matrix of the resulting estimator of α_m, is given by $F = \Sigma_{mm}^{-1}$. In other words, any other transformation would lead to an estimator with an asymptotic variance–covariance matrix 'greater' than the one obtained using Σ_{mm}^{-1} ('greater' is used to mean that the difference would be positive definite). This result is based on Theorem 5 of Balestra (1983). Its application to our model can be found in Krishnakumar (1988).

Therefore the optimal instrument for Z_m is given by $\Sigma_{mm}^{-1}X$ and premultiplying (10.56) by $X'\Sigma_{mm}^{-1}$, we get:

$$X'\Sigma_{mm}^{-1}y_m = X'\Sigma_{mm}^{-1}Z_m\alpha_m + X'\Sigma_{mm}^{-1}u_m \ . \tag{10.57}$$

Applying GLS on (10.57), we obtain what we call the generalised two stage least squares (G2SLS) estimator of α_m:

$$\begin{aligned}\widehat{\alpha}_{m,G2\text{SLS}} =& [Z_m'\Sigma_{mm}^{-1}X(X'\Sigma_{mm}^{-1}X)^{-1}X'\Sigma_{mm}^{-1}Z_m]^{-1} \\ & \times Z_m'\Sigma_{mm}^{-1}X(X'\Sigma_{mm}^{-1}X)^{-1}X'\Sigma_{mm}^{-1}y_m \end{aligned} \tag{10.58}$$

Now, the above estimator is not feasible as Σ_{mm} is unknown. Hence we need a prior estimation of the variance components. By analysis of variance (cf. Amemiya (1971)) of the errors of the m–th structural equation, the following estimators of the σ_{imm}(s) are obtained:

$$\begin{aligned}\widetilde{\sigma}_{1mm} &= \frac{1}{(N-1)(T-1)}u_m'M_1u_m \\ \widetilde{\sigma}_{2mm} &= \frac{1}{N-1}u_m'M_2u_m \\ \widetilde{\sigma}_{3mm} &= \frac{1}{T-1}u_m'M_3u_m \\ \widetilde{\sigma}_{4mm} &= \widetilde{\sigma}_{2mm} + \widetilde{\sigma}_{3mm} - \widetilde{\sigma}_{1mm} \end{aligned} \tag{10.59}$$

These formulae contain u_m which is also unknown. However, it can be estimated as follows. Premultiplying (10.56) by the instrument M_1X, we get:

$$X'M_1y_m = X'M_1Z_m\alpha_m + X'M_1u_m \ .$$

Note that, if the equation has an intercept, it gets eliminated by this transformation and we will be left with:

$$X'M_1y_m = X'M_1Z_m^*\alpha_m^* + X'M_1u_m \tag{10.60}$$

where Z_m^* denotes the matrix of right hand side variables excluding the vector of ones and α_m^* the respective coefficients. That is, we have split Z_m and α_m as:

$$Z_m = [Y_m l_{NT} X_m^*]\,; \quad \alpha_m = \begin{pmatrix} \gamma_m \\ a_m \\ b_m \end{pmatrix}$$

and redefined Z_m^* and α_m^* as

$$Z_m^* = [Y_m\ X_m^*]\,; \quad \alpha_m^* = \begin{pmatrix} \gamma_m \\ b_m \end{pmatrix}\,.$$

Performing GLS on (10.60), we obtain a consistent estimator of α_m^* called the covariance or the within 2SLS estimator:

$$\widehat{\alpha}^*_{m,\text{cov2SLS}} = [Z_m^{*'} M_1 X (X' M_1 X)^{-1} X' M_1 Z_m^*]^{-1} Z_m^{*'} M_1 X (X' M_1 X)^{-1} X' M_1 y_m \quad (10.61)$$

The intercept is estimated as:

$$\widehat{a}_{m,\text{cov2SLS}} = \frac{1}{NT} l'_{NT} (y_m - Z_m^* \widehat{\alpha}^*_{m,\text{cov2SLS}})\,.$$

From these estimators, we can predict u_m as:

$$\widehat{u}_{m,\text{cov2SLS}} = y_m - Z_m^* \widehat{\alpha}_{m,\text{cov2SLS}} - l_{NT}\, \widehat{a}_{m,\text{cov2SLS}}\,.$$

Substituting $\widehat{u}_{m,\text{cov2SLS}}$ for u_m in (10.59), we obtain $\widehat{\sigma}_{imm},\ i = 1,\ 2,\ 3,\ 4$ and $\widehat{\Sigma}_{mm} = \sum_{i=1}^{4} \widehat{\sigma}_{imm} M_i$, leading to the following feasible G2SLS estimator of α_m:

$$\begin{aligned} \widehat{\alpha}_{m,\text{fG2SLS}} = {} & [Z_m' \widehat{\Sigma}^{-1}_{mm} X (X' \widehat{\Sigma}^{-1}_{mm} X)^{-1} X' \widehat{\Sigma}^{-1}_{mm} Z_m]^{-1} \\ & \times Z_m' \widehat{\Sigma}^{-1}_{mm} X (X' \widehat{\Sigma}^{-1}_{mm} X)^{-1} X' \widehat{\Sigma}^{-1}_{mm} y_m \end{aligned} \quad (10.62)$$

Before giving the limiting distribution of the above estimators, we just mention that all our estimators are consistent. Another interesting point to note is that all the three estimators—Cov2SLS, G2SLS, fG2SLS—have the same limiting distribution. It is given by (see Krishnakumar (1988) for derivation):

$$\begin{pmatrix} \sqrt{N}\ (\widehat{a}_m - a_m) \\ \sqrt{NT}\ (\widehat{\alpha}_m^* - \alpha_m^*) \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_{\mu mm} + \sigma_{\varepsilon mm} & 0 \\ 0 & \sigma_{\nu mm} (\widetilde{P}_m' \widetilde{R}_m \widetilde{P}_m)^{-1} \end{pmatrix}\right)$$

where

$$\widetilde{P}_m = \begin{pmatrix} 1 & 0 \\ 0 & \Pi_{*m} \\ 0 & H_m^* \end{pmatrix}\ ; \quad \widetilde{R}_m = \begin{pmatrix} 1/(\sigma_{\mu mm} + \sigma_{\varepsilon mm}) & 0 \\ 0 & 1/(\sigma_{\nu mm}) R \end{pmatrix}$$

with Π_{*m} being the coefficient matrix of X in the reduced form equations for Y_m except for the column of ones and H_m^* being a selection matrix such that $X_m^* = X H_m^*$.

Generalised Three Stage Least Squares

The extension from G2SLS to generalised 3SLS (G3SLS) can be done in two ways. In what follows, we present both the ways and show that they yield asymptotically equivalent estimators.

The reader will recall that the G2SLS method uses the instrument matrix $\Sigma_{mm}^{-1}X$ for the m–th equation. Applying to each structural equation of the system, its corresponding transformation given by $\Sigma_{mm}^{-1}X$, $m = 1,\ 2,\ \ldots\ ,M$, we obtain:

$$\begin{aligned} X'\Sigma_{11}^{-1}y_1 &= X'\Sigma_{11}^{-1}Z_1\alpha_1 + X'\Sigma_{11}^{-1}u_1 \\ &\vdots \\ X'\Sigma_{MM}^{-1}y_M &= X'\Sigma_{MM}^{-1}Z_M\alpha_M + X'\Sigma_{MM}^{-1}u_m \end{aligned}$$

or

$$\widetilde{X}'^{-1}y = \widetilde{X}'^{-1}Z\alpha + \widetilde{X}'^{-1}u \tag{10.63}$$

where[12]

$$\begin{aligned} \widetilde{X} &= I \otimes X \\ D &= \text{ diag } [\Sigma_{11}\ \ldots\ \Sigma_{MM}] \\ Z &= \text{ diag } [Z_1\ \ldots\ Z_M] \\ \alpha' &= [\alpha_1'\ \ldots \alpha_M'] \\ u' &= [u_1' \ldots u_M'] \\ y' &= [y_1' \ldots y_M'] \end{aligned}$$

Now, let us apply GLS to the transformed system (10.63) to obtain our first generalised 3SLS (G3SLS–I) estimator:

$$\begin{aligned} \widehat{\alpha}_{\text{G3SLS–I}} = &[Z'D^{-1}\widetilde{X}(\widetilde{X}'D^{-1}\Sigma D^{-1}\widetilde{X})^{-1}\widetilde{X}D^{-1}Z]^{-1} \\ &\times Z'D^{-1}\widetilde{X}(\widetilde{X}'D^{-1}\Sigma D^{-1}\widetilde{X})^{-1}\widetilde{X}'D^{-1}y \end{aligned} \tag{10.64}$$

Note that this way of generalising is analogous to the way that classical 2SLS is extended to 3SLS by Zellner and Theil (1962). However, there is also a second way of approaching the problem, that we briefly present below.

Recall that our reason for choosing $\Sigma_{mm}^{-1}X$ as the instrument for Z_m in the G2SLS procedure was that it minimised the asymptotic covariance matrix of the resulting coefficient estimator. Now, let us write the whole system as:

$$y = Z\alpha + u$$

with

$$E(u) = 0 \quad \text{and} \quad E(uu') = \Sigma$$

[12] Note that this particular notation for D is valid only for this chapter.

and find the best transformation F of $(I \otimes X)$ for choosing the instruments. By the same reasoning as for G2SLS, we would get $F = \Sigma^{-1}$. Using $\Sigma^{-1}(I \otimes X)$ as instruments and estimating α by GLS on the transformed system yields our second G3SLS (G3SLS–II) estimator:

$$\begin{aligned}\widehat{\alpha}_{\text{G3SLS}-\text{II}} = &[Z'\Sigma^{-1}\widetilde{X}(\widetilde{X}'\Sigma^{-1}\widetilde{X})^{-1}\widetilde{X}'\Sigma^{-1}Z]^{-1} \\ &\times Z'\Sigma^{-1}\widetilde{X}(\widetilde{X}'\Sigma^{-1}\widetilde{X})^{-1}\widetilde{X}'\Sigma^{-1}y \end{aligned} \tag{10.65}$$

Both these G3SLS estimators can be made feasible by replacing the variance components present in Σ by their corresponding estimates given by analysis of variance:

$$\begin{aligned}\widehat{\sigma}_{1mm'} &= \frac{1}{(N-1)(T-1)} \quad \widehat{u}'_m M_1 \widehat{u}_{m'} \\ \widehat{\sigma}_{2mm'} &= \frac{1}{N-1} \quad \widehat{u}'_m M_2 \widehat{u}_{m'} \\ \widehat{\sigma}_{3mm'} &= \frac{1}{T-1} \quad \widehat{u}'_m M_3 \widehat{u}_{m'} \\ \widehat{\sigma}_{4mm'} &= \widehat{\sigma}_{2mm'} + \widehat{\sigma}_{3mm'} - \widehat{\sigma}_{1mm'} \end{aligned} \tag{10.66}$$

for $m,\ m' = 1,\ \ldots\ M$. Note that for the $\widehat{u}_m$(s), we can take:

$$\widehat{u}_{m,\text{cov2SLS}} = y_m - Z_m \widehat{\alpha}_{m,\text{cov2SLS}}$$

or

$$\widehat{u}_{m,\text{fG2SLS}} = y_m - Z_m \widehat{\alpha}_{m,\text{fG2SLS}}$$

or even

$$\widehat{u}_{m,\text{cov3SLS}} = y_m - Z_m \widehat{\alpha}_{m,\text{cov3SLS}}$$

where $\widehat{\alpha}_{m,\text{cov3SLS}}$ is yet another 3SLS estimator obtained by using the instrument matrix $(I \otimes M_1 X)$ for the system and estimating by GLS.

From the estimates of $\widehat{\sigma}_{imm'}$(s) given by (10.66), we form $\widehat{\Sigma} = \sum_{i=1}^{4} \widehat{\Sigma}_i \otimes M_i$ with $\widehat{\Sigma}_i = [\widehat{\sigma}_{imm'}],\ m, m' = 1,\ \ldots,\ M$ and use it in (10.64) and (10.65) to get the feasible G3SLS estimators.

It is remarkable that due to the special structure of the error-components covariance matrix, all these 3SLS estimators, namely the pure G3SLS-I, pure G3SLS-II, cov3SLS, feasible G3SLS-I and feasible G3SLS-II, have the same limiting distribution given by:

$$\begin{pmatrix} \sqrt{N}\ (\widehat{a} - a) \\ \sqrt{NT}\ (\widehat{\alpha}^* - \alpha^*) \end{pmatrix} \ \sim\ N\left(0, \begin{pmatrix} \Sigma_\mu + \Sigma_\varepsilon & 0 \\ 0 & [\bar{\Pi}'(\Sigma_v^{-1} \otimes R)\bar{\Pi}]^{-1} \end{pmatrix}\right)$$

where a is a $(M \times 1)$ column vector containing the intercepts of each equation i.e. $a' = [a_1\ \ldots\ a_M]$ and α^* is $((M-1)M \times 1)$ containing the other non-zero

coefficients of each equation i.e. $\alpha^{*'} = [\alpha_1^{*'} \ \ldots \ \alpha_M^{*'}]$ and where $\bar{\Pi} = \text{diag}\,([\Pi_m^* H_m^*])$, $m = 1, \ldots, M$.

Finally, let us note that, though we assume the presence of an intercept in each equation, the above results can be easily generalised to the case in which some equations have an intercept and others do not.

Error Components Two Stage Least Squares

This is an alternative method of estimating the parameters of a single structural equation. This method is proposed by Baltagi (1981) and inspired from the feasible Aitken procedure developed by Maddala (1971) for a single equation error components model.

In this method, the structural equation in question say the m–th one, is successively transformed by the matrices of eigenvectors associated with the distinct characteristic roots of Σ_{mm} and GLS is performed on a system comprising all the three transformed equations. Before going further, let us introduce some more notations. From Sect. 10.3.1.1 we know that the distinct eigenvalues of $\Sigma_{mm'}$ are $\sigma_{1mm'}$, $\sigma_{2mm'}$, $\sigma_{3mm'}$ and $\sigma_{4mm'}$. The matrices whose columns are the eigenvectors associated with these roots are Q_1 , Q_2 , Q_3 and $l_{NT}/\sqrt{NT}$ respectively where $Q_1 = C_2 \otimes C_1$, $Q_2 = C_2 \otimes l_T/\sqrt{T}$, $Q_3 = l_N/\sqrt{N} \otimes C_1$ such that $O_T' = [l_T'/\sqrt{T} \quad C_1']$ and $O_N' = [l_N'/\sqrt{N} \quad C_2']$ are orthogonal. Note that $Q_j Q_j'$ are unique for $j = 1,\ 2,\ 3$ and $Q_j' Q_j = M_j,\ j = 1,\ 2,\ 3$.

Now, let us apply the transformations $Q_j,\ j = 1,\ 2,\ 3$ to our structural equation (10.56):

$$Q_j y_m = Q_j Z_m \alpha_m + Q_j u_m, \quad j = 1,\ 2,\ 3\ . \tag{10.67}$$

It is easily verified that

$$E(Q_j u_m u_m' Q_j') = \begin{cases} \sigma_{jmm} I_{m_j} & \text{for} \quad j = j' \\ 0 & \text{for} \quad j \neq j' \end{cases}$$

Thus the transformed errors have a scalar variance–covariance matrix but are still correlated with the right hand side variables. Hence an IV technique is used with $Q_j X$ as instruments for $Q_j Z_m$. This gives:

$$\begin{aligned} \widehat{\alpha}_{m,2\text{SLS}}^{(j)} = & [Z_m' Q_j' Q_j X (X' Q_j' Q_j X)^{-1} X' Q_j' Q_j Z_m]^{-1} \\ & \times Z_m' Q_j' Q_j X (X' Q_j' Q_j X)^{-1}\ X' Q_j' Q_j y_m, \quad j = 1,\ 2,\ 3 \end{aligned} \tag{10.68}$$

These 2SLS estimators are in turn used to estimate the variance components:

$$\begin{aligned} \widehat{\sigma}_{jmm} &= \frac{1}{m_j} (Q_j y_m - Q_j Z_m \widehat{\alpha}_{m,2\text{SLS}}^{(j)})' (Q_j y_m - Q_j Z_m \widehat{\alpha}_{m,2\text{SLS}}^{(j)}), \quad j = 1,\ 2,\ 3 \\ \widehat{\sigma}_{4mm} &= \widehat{\sigma}_{2mm} + \widehat{\sigma}_{3mm} - \widehat{\sigma}_{1mm} \end{aligned} \tag{10.69}$$

This is a generalisation of the Swamy and Arora (1972) method. The above procedure gives three different estimators of the same α_m. Therefore, we can combine all the three transformed equations of (10.67) together and estimate the whole system by GLS. We have:

$$\begin{pmatrix} X'Q_1'Q_1y_m \\ X'Q_2'Q_2y_m \\ X'Q_3'Q_3y_m \end{pmatrix} = \begin{pmatrix} X'Q_1'Q_1Z_m \\ X'Q_2'Q_2Z_m \\ X'Q_3'Q_3Z_m \end{pmatrix} \alpha_m + \begin{pmatrix} X'Q_1'Q_1u_m \\ X'Q_2'Q_2u_m \\ X'Q_3'Q_3u_m \end{pmatrix} . \tag{10.70}$$

Using the Swamy and Arora estimates (10.69) of the variance components and performing feasible GLS on (10.70), we get the error components two stage least squares (EC2SLS) estimator:

$$\widehat{\alpha}_{m,\text{EC2SLS}} = \left[\sum_{j=1}^{3} \frac{1}{\widehat{\sigma}_{jmm}} Z_m'Q_j'Q_jX(X'Q_j'Q_jX)^{-1}X'Q_j'Q_jZ_m\right]^{-1} \tag{10.71}$$

$$\times \left[\sum_{j=1}^{3} \frac{1}{\widehat{\sigma}_{jmm}} [Z_m'Q_j'Q_jX(X'Q_j'Q_jX)^{-1}X'Q_j'Q_jy_m]\right] \tag{10.72}$$

It can be shown that the above estimator is a weighted average of the three 2SLS estimators given in (10.68).

The limiting distribution of the EC2SLS estimator is the same as that of the feasible G2SLS estimator.

Error Components Three Stage Least Squares

In this section, we present an extension of the EC2SLS method to the whole system. We start with

$$y = Z\alpha + u$$

and transform it successively by $(I_M \otimes Q_j)$, $\quad j = 1,\ 2,\ 3$ to give:

$$y^{(j)} = Z^{(j)}\alpha + u^{(j)}, \quad j = 1,\ 2,\ 3 \tag{10.73}$$

where $y^{(j)} = (I_M \otimes Q_j)y$; $Z^{(j)} = (I_M \otimes Q_j)Z$; $u^{(j)} = (I_M \otimes Q_j)u$ $\quad$ and

$$E(u^{(j)}u^{(j)'}) = \Sigma_j \otimes I_{mj}, \quad j = 1,\ 2,\ 3\ .$$

Using $X^{(j)} = (I_M \otimes Q_jX)$ as instruments for $Z^{(j)}$ and applying GLS, we get:

$$\widehat{\alpha}_{\text{IVGLS}}^{(j)} = [Z^{(j)'}\{\Sigma_j^{-1} \otimes P_{X^{(j)}}\}Z^{(j)}]^{-1} \tag{10.74}$$

$$\times [Z^{(j)'}\{\Sigma_j^{-1} \otimes P_{X^{(j)}}\}y^{(j)}] \quad j = 1,\ 2,\ 3$$

where for any matrix A, P_A denotes the projection matrix $A(A'A)^{-1}A'$.

The unknown variance components are estimated by

$$\widehat{\sigma}_{jmm'} = \frac{1}{m_j}[Q_j y_m - Q_j Z_m \widehat{\alpha}^{(j)}_{m,2\text{SLS}}]' \, [Q_j y_{m'} - Q_j Z_{m'} \widehat{\alpha}^{(j)}_{m',2\text{SLS}}] \,, \quad j = 1,\ 2,\ 3$$

$$\widehat{\sigma}_{4mm'} = \widehat{\sigma}_{2mm'} + \widehat{\sigma}_{3mm'} - \widehat{\sigma}_{1mm'}$$

Now, recognising once again that the same α is being estimated three times separately, we can combine all the three transformed systems of (10.73) and estimate the global system by (feasible) IVGLS. The resulting estimator is called the error component 3SLS (EC3SLS) estimator of α:

$$\widehat{\alpha}_{\text{EC3SLS}} = \left[\sum_{j=1}^{3} Z^{(j)'}(\widehat{\Sigma}_j^{-1} \otimes P_{X^{(j)}})Z^{(j)}\right]^{-1} \tag{10.75}$$

$$\times \left[\sum_{j=1}^{3} Z^{(j)'}(\widehat{\Sigma}_j^{-1} \otimes P_{X^{(j)}})y^{(j)}\right] \tag{10.76}$$

The above estimator also has the same interpretation as the EC2SLS one, in that it is a weighted average of the three 3SLS estimators of (10.75) (see Baltagi (1981) for further details).

Finally, the limiting distribution of the EC3SLS estimator can be shown to be the same as that of the G3SLS estimators of the previous section and hence is asymptotically equivalent to them.

Full Information Maximum Likelihood

The full information maximum likelihood (FIML) procedure consists in maximising the log-likelihood function of the model with respect to the structural parameters given the *a priori* restrictions. As in all constrained maximisation problems, there are two ways of tackling it—(i) by maximising the corresponding Lagrangian function with respect to the same parameters and a set of multipliers associated with the constraints; (ii) by substituting the constraints in the objective function and performing maximisation without constraints. In this section, we will briefly review both the approaches. The reader will note that neither of them yield explicit analytical solutions and hence both require numerical iterative procedures to arrive at the solution. Moreover, in the first approach adopted by Balestra and Krishnakumar (1987) and Krishnakumar (1988), the *a priori* restrictions on the structural coefficients are assumed to be any linear ones whereas in the second approach followed by Prucha (1985), only the normalisation and exclusion restrictions are considered.

Recalling our structural model:

$$Y\Gamma + XB + U = 0.$$

and separating the intercept of each equation from the other terms, we can write:

$$Y\Gamma + l_{NT}a' + X_*B_* + U = 0$$

or

$$l_{NT}a' + Z_*\Theta_* + U = 0$$

where

$$\Theta_* = \begin{pmatrix} \Gamma \\ B_* \end{pmatrix}.$$

Note that in case only some equations have an intercept and others do not, the following procedure can be easily modified accordingly.

Now, the a priori restrictions on the coefficients can be written as (say we have p of them):

$$\begin{pmatrix} S_0 & 0 \\ 0 & S_* \end{pmatrix} \begin{pmatrix} a \\ \text{vec}\,\Theta_* \end{pmatrix} = \begin{pmatrix} s_0 \\ s_* \end{pmatrix} \tag{10.77}$$

These include the normalisation rule, the exclusion restrictions and any other linear constraints. To these, we add the symmetry conditions for Σ_j(s) written as:

$$C \;\; \text{vec} \;\; \Sigma_j = 0\,, \quad j = \mu\,, \varepsilon\,, \nu\,. \tag{10.78}$$

The log-likelihood function of the model can be written as follows, after a few simplifications and rearrangements:

$$\begin{aligned} lnL = \text{const} \; &- \frac{1}{2}\sum_{i=1}^{4} m_i ln \mid \Sigma_i \mid + \frac{1}{2} NT ln \mid L'\Theta_* \mid^2 \\ &- \frac{1}{2} tr(NTaa' + \Theta_*'Z_*'l_{NT}a' + al_{NT}'Z_*\Theta_*)\Sigma_4^{-1} \\ &- \frac{1}{2} tr \sum_{i=1}^{4} \Theta_*'Z_*'M_iZ_*\Theta_*\Sigma_i^{-1} \end{aligned} \tag{10.79}$$

with L such that $\Gamma = L\Theta_*$. Thus we have to maximise (10.79) with respect to a, Θ_*, Σ_μ, Σ_ε and Σ_ν under the constraints (10.77) and (10.78).

Here again we will not describe the procedure in detail for brevity's sake and the reader is invited to consult Balestra and Krishnakumar (1987) for more information on the algorithm to be implemented in order to obtain a numerical solution. We give below the limiting distribution of the FIML estimator:

$$\begin{pmatrix} \sqrt{T} & (\widehat{a}_{ML} - a) \\ \sqrt{NT} & \text{vec}(\widehat{\Theta}_{*,ML} - \Theta_*) \end{pmatrix} \sim N\left(0, \begin{pmatrix} \Sigma_\mu + \Sigma_\varepsilon & 0 \\ 0 & F[F'(\Sigma_\nu^{-1} \otimes P_*)F]^{-1}F' \end{pmatrix}\right)$$

where

$$P_* = \begin{pmatrix} \Pi_*' \\ I \end{pmatrix} R \quad (\Pi_* \;\; I)$$

When the a priori restrictions are only the normalisation and exclusions, we have the same limiting distribution for the FIML as the one for the (feasible) G3SLS. Hence, in this case, the FIML and the fG3SLS are of the same asymptotic efficiency.

As mentioned in the beginning of this section, there is a second approach to the constrained maximisation problem which consists in replacing the constraints in the objective function and then maximising the latter with no constraints. This has been done by Prucha (1985) for our model for the case of the usual restrictions only and is called the normal FIML (NFIML) estimator. The normal equations of the above maximisation programme lead to an IV interpretation of the ML estimator which can be used as an estimator–generating equation to form a general class of estimators called the NFIML$_A$ estimator (the subscript A indicates that the estimator can be viewed as an approximation of the NFIML estimator). Further Prucha also shows that under certain conditions, all members of the NFIML$_A$ class are asymptotically equivalent among themselves and to the NFIML estimator.

10.3.1.4 Asymptotic Comparisons of the Various Structural Estimators

In this section, we will summarise the different asymptotic equivalences mentioned earlier and state a few more results regarding the just–identified case.

First, let us briefly recall the results that we already know in the case of the usual restrictions. We have the asymptotic equivalence of the various 2SLS estimators namely, the cov2SLS, fG2SLS and EC2SLS. Among the system methods, we have the asymptotic equivalence of cov3SLS, fG3SLS-I, fG3SLS-II, EC3SLS and FIML estimators.

Regarding the just-identified case, we will mention the important results without deriving them. The reader is referred to the original works by Krishnakumar (1988) and Baltagi (1981) for proofs.

When a single equation, say the m-th one, is just-identified:

(i) the indirect least squares estimators obtained using the covariance estimator of Π, is exactly equal to the cov2SLS estimator;
(ii) the indirect least squares estimator obtained using the feasible GLS estimator of Π has the same limiting distribution as the feasible G2SLS estimator;
(iii) the EC2SLS estimator can be expressed as a weighted combination of three indirect estimators of α_m;
(iv) the three 2SLS estimators of (10.68) are respectively equal to the indirect least squares estimators based on the between groups, between time periods and within variations estimators of the reduced form;
(v) all these estimators—feasible G2SLS, cov2SLS, indirect estimators based on $\widehat{\Pi}_{\text{cov}}$ or $\widehat{\Pi}_{\text{fGLS}}$ and EC2SLS—are asymptotically equivalent.

When the whole system is just-identified:

(i) fG3SLS-I reduces to fG2SLS whereas the fG3SLS-II does not;
(ii) fG3SLS-I, fG3SLS-II, fG2SLS and the indirect estimators are all asymptotically equivalent;
(iii) EC3SLS does not reduce to EC2SLS;

(iv) EC3SLS and EC2SLS have the same limiting distribution and
(v) all these estimators—fG3SLS-I, fG3SLS-II, cov3SLS, fG2SLS, cov2SLS, EC3SLS, EC2SLS—are asymptotically equivalent.

10.3.1.5 Small Sample Properties

There are essentially two ways of arriving at the small sample behaviour of econometric estimators. One is by analytically deriving the exact distribution or an approximation to it and the other is by 'constructing' the distribution through simulations (also called Monte-Carlo experiments).

In the case of the reduced form, the unbiasedness of the various coefficient and variance components estimators is proved without great difficulty (see Krishnakumar (1988)). However, exact efficiency properties are yet to be established and so far nothing is known.

In the case of the structural form estimators, things get very complicated. In the classical simultaneous model, several authors have dealt with the problem of finding the exact distributions of the two stage and three stage estimators. The reader is invited to consult Phillips (1982) for more information in the classical case. In the SEM with EC case, we have no result on the exact density functions of the various structural estimators. However, we do have results on approximations to finite sample moments using series expansions methods. These methods are used even when we have the analytical expression of density functions since they yield much less complicated expressions. In these methods, the estimator is developed around its true value in a series of terms of orders decreasing in the powers of the sample size. Then the series is truncated upto a desired order and the expectation of the truncated series is calculated to get the bias upto that order. This procedure has been applied to our model by Krishnakumar (1988), following the approach of Nagar (1959), to get approximations for the bias of cov2SLS and fG2SLS estimators. We will not go deeper into this aspect here. The results and derivations can be found in Krishnakumar (1988).

Now, we turn to the second approach—the Monte-Carlo study. This method consists in specifying a true model giving values for all the parameters, generating the random elements and the observations on the exogenous variables, calculating the endogenous variables and estimating the parameters using only the observations. By running the procedure a number of times with different sets of observations (keeping the true values unchanged), one can 'construct' the distribution curve of the estimator and derive its mean, variance, mean square error and so on. These criteria can be used to compare the performance of different estimation methods. In addition, the whole exercice can be repeated for different sets of true values.

Baltagi (1984) carried out such a Monte-Carlo experiment for the SEM with EC, in which he compared various least squares and IV estimators of a two-equation structural model, keeping the same true values for the coefficients and changing only the values of the variance components. In what follows, we will briefly review

the main results concerning the structural form and the reduced form. For results regarding the variance components estimators, the reader is referred to Baltagi (1984).

First, the structural form results. The classical 2SLS has a smaller bias than the EC2SLS but the EC2SLS has a lower root mean square error (RMSE) than the classical 2SLS. Better estimates of the structural variance components do not necessarily imply better estimates of the structural coefficients. In general, 3SLS dominates 2SLS and EC3SLS dominates EC2SLS in RMSE though the superiority of EC3SLS over EC2SLS does not hold for all the structural parameters. There is gain in performing EC3SLS rather than classical 3SLS according to RMSE. Similar results are also obtained if we use global criteria like the normalised mean square deviation and the normalised mean absolute deviation which give a single indicator for the combined performance of all parameter estimators.

Now, the reduced form results. Performing feasible GLS on each reduced form equation is better than performing OLS or LSDV, according to RMSE. But, according to the same criterion, feasible GLS on the entire system does not necessarily produce better results than feasible GLS on each equation separately. Baltagi notes that this could be due to the fact that there are only two equations in the model and may not be so in larger models. Once again better estimates of the variance components do not necessarily imply better feasible GLS estimates of coefficients. The same results are maintained even according to global criteria.

Mátyás, and Lovrics (1990) investigate the small scale properties of 5 limited information estimators for SEM with EC models by means of a Monte Carlo study. They compare the OLS estimator, the within estimator, the pure G2SLS, and two feasible G2SLS estimators (one with OLS as the first step and the other with within). Their findings are as follows: The OLS estimator remains biased in all cases. But it is still recommended for very small N and T ($N < 10,\ T < 20$) due to its stability as the G2SLS/within 2SLS are unstable and have a large dispersion. For $N < 10$ and $T > 20$ they favour the G2SLS/within 2SLS estimators and for a sufficient ($N > 15 - 20$) as long as $T > 5$. There is practically no difference between the three G2SLS (pure and the two feasible) estimators.

Baltagi and Chang (2000) study the relative performance of several estimators of a two-equation SEM with unbalanced panel data. Among the single equation methods they compare 2SLS, W2SLS and the EC2SLS and among the system estimators they look at 3SLS, W3SLS and EC3SLS. They observe that most of the results obtained for the balanced case carry over to the unbalanced one.

10.3.2 Extensions

10.3.2.1 Simultaneous Equation Models with Correlated Specific Effects

In the SEM with EC discussed in the previous subsections, it was assumed that the error components were uncorrelated with the exogenous variables. Cornwell et al. (1992) extend our model to the case in which this assumption is dropped.

They allow for the possibility of only some variables being correlated with the error components (singly exogenous) while the others are independent of them (doubly exogenous). Their model is specified as follows:

$$y_m = Y_m\delta_m + X_m\beta_m + Z_m\gamma_m + \alpha_m + \varepsilon_m \,, m = 1, \ldots, M \,. \tag{10.80}$$

A distinction is also made between time-varying exogenous variables (X) and the time-invariant exogenous variables (Z) and only individual effects are present in the model (i.e. we have only a two-components error term). Denoting,

$$R_m = [Y_m\ X_m\ Z_m]\,; \quad \underline{\xi}'_m = [\underline{\delta}'_m\ \underline{\beta}'_m\ \underline{\gamma}'_m]\,,$$

we can write (10.80) as

$$y_m = R_m\xi_m + (\alpha_m + \varepsilon_m)\,, \quad m = 1, \ldots, M$$

The 2SLS proposed transforms the equation say the first one by $\Sigma_{11}^{-\frac{1}{2}}$ to get:

$$\Sigma_{11}^{-\frac{1}{2}} y_1 = \Sigma_{11}^{-\frac{1}{2}}\ R_1\ \xi_1 + \Sigma_{11}^{-\frac{1}{2}}(\alpha_1 + \varepsilon_1)$$

and use instruments of the form $A = [Q_v X \quad P_v B]$ with different choices for B. Three different choices are proposed the first one corresponding to the instrument set of Hausman and Taylor (1981), the second one inspired from Amemiya and McCurdy (1986) and the third based on Breusch (1987). The Three Stage Least Squares generalises the procedure for the whole model.

The authors also derive estimators in the case in which the nature of the correlation between the exogenous variables and the specific effects may vary from equation to equation. In other words, we may have an exogenous variable correlated with the specific effect in one equation but uncorrelated with the specific effect in another equation. In this case, the instrument set also varies across equations.

In case the specific effects are assumed to be fixed the authors show that the model can be estimated by OLS after a within transformation.

10.3.2.2 Simultaneous Error Component Models with Censored Endogenous Variables

Another recent extension is the inclusion of censored endogenous variables in a simultaneous EC model, by Vella and Verbeek (1999). Their model is a two–equation system in which the first one is the primary focus and the second one is already in the reduced form. For $i = 1, \ldots, N;\ t = 1, \ldots, T$ we have:

$$y^*_{it} = m_1(x_{it}, z_{it}; \theta_1) + \mu_i + \eta_{it} \tag{10.81}$$

$$z^*_{it} = m_2(x_{it}, z_{it}; \theta_2) + \alpha_i + \nu_{it} \tag{10.82}$$

$$z_{it} = h(z^*_{it}, \theta_3)$$
$$y_{it} = k(y^*_{it})$$

where i indexes individuals ($i = 1,\ldots,N$), t time periods ($t = 1,\ldots,T$), y^*_{it} and z^*_{it} are latent endogenous variables with observed counterparts y_{it} and z_{it}; m_1 and m_2 denote general functions characterized by the unknown parameters in θ_1 and θ_2, respectively. The mapping from the latent to the observed variables is through the censoring functions h and k, h depending on another unknown parameter vector θ_3.

An error component structure is specified for the disturbance term of each equation (μ_i and η_{it} for (10.81) and α_i and ν_{it} for (10.82)) with the components being independent across individuals. Denoting $\varepsilon_{it} = \mu_i + \eta_{it}$ and $u_{it} = \alpha_i + \nu_{it}$, it is assumed that

$$u_i \mid X_i \sim NID(0, \sigma^2_\alpha \iota \iota'^2_\nu I),$$
$$E(\varepsilon_{it} \mid X_i, u_i) = \tau_1 u_{it} + \tau_2 \bar{u}_i \tag{10.83}$$

where ι is a vector of ones, u_i is the T vector of u_{it}s for individual i, $X_i = [x_{i1},\ldots,x_{iT}]'$ and $\bar{u}_i = T^{-1}\sum_{t=1}^{T} u_{it}$; τ_1 and τ_2 are unknown constants. Equation (10.83) reflects the endogenous character of z^*_{it}.

Two variants are considered for the censoring mechanisms:
(1) z^*_{it} is censored through $h(\cdot)$ and y^*_{it} observed only for certain values of $z_{i1},\ldots,z_{iT}$ i.e

$$y_{it} = y^*_{it} \quad \text{if} \quad g_t(z_{i1},\ldots,z_{it}) = 1$$
$$= 0 \quad \text{(unobserved)} \quad \text{if} \quad g_t(z_{i1},\ldots,z_{iT}) = 0$$

and
(2) z^*_{it} is observed and only y^*_{it} is censored through $k(\cdot)$.

The first model allows for a conditional moment estimation where (10.82) is first estimated by ML and then (10.81) by conditional moment method after adding in its right hand side the conditional expectation of its errors given the exogenous variables and the errors of (10.82), in order to take into account the endogeneity of z_{it}. For the second variant, a two step conditional ML approach is proposed by first estimating the second equation by ML as z_{it} is observed and then the first equation by conditional ML i.e. maximising the conditional likelihood given z_i. Generalisations to multiple endogenous variables are briefly mentioned.

The first method is applied to a model for analyzing the influence of the number of hours worked on the hourly wage rate keeping in mind the potential endogeneity of the former. Through this application the authors point out the usefulness of the two step methods in a context where the maximum likelihood procedure is impractical.

10.4 Conclusion

To conclude we would like to make a few general remarks. First let us add a word on the different uses of the same terminology and a possible confusion arising from it, especially for students. As mentioned before, the problem of regressors correlated with the error term (whatever component of it) results in inconsistent/biased OLS/GLS estimates and one has to resort to IV/GMM methods. When data are in a one-dimensional form, there is no room for confusion. However in a panel data setting, the same terminology of 'endogeneity of regressors' may be used whether it concerns correlation with specific effects or with the residual disturbance term. Though it is correct to use the same name in both cases, the researcher has to check what type of endogeneity she is faced with before adopting a solution. Some methods or transformations that are valid for one may not be valid for the other and vice versa.

Again keeping the students in mind we would like to point out that the terms IV and GMM can rightly be used in an interchangeable fashion as all IV estimators can also be interpreted as GMM estimators using the corresponding moment conditions. But one should understand how the same estimator can be obtained by both ways especially for implementing the estimation methods in any software package which may not explicitly have one or the other term in its commands.

We now turn to areas where research could be continued in this topic. First of all, the reader would have noticed that we have not specially dealt with hypothesis testing in our chapter. This is because the tests on various coefficients and variance components are only asymptotic, based on the limiting distributions of the respective estimators and can be derived relatively easily as straightforward extensions of their counterparts in the single-equation model. No exact results are available so far on the distributions. This precisely leads us to one possible area for further theoretical research namely, derivation of the exact distributions of the various estimators developed above, or better approximations to the exact distribution than the asymptotic ones, especially for small samples, using recent techniques like bootstrap or saddlepoint approximations.

Finally, regarding the practical implementation of the various IV methods, we are happy to note that many of the above procedures have been included in the econometric software available on the market. G2SLS, within-2SLS and EC-2SLS are easily implemented in STATA which offers many estimation and inference possibilities with panel data in general. Matrix manipulations are also convenient in this programme which allows for easy and quick transformations of variables before entering them in a regression. Other packages like TSP, LIMDEP and RATS have also included panel data estimation possibilities. The reader is invited to go through the chapter devoted to this topic in this volume for an excellent review of the different options available.

References

Ahn, S.C., and P. Schmidt (1995): Efficient Estimation of Models for Dynamic Panel Data. *Journal of Econometrics*, 68, 5–27.

Amemiya, T. (1971): The Estimation of Variances in a Variance Components Model. *International Economic Review,* 12, 1–13.

Amemiya, T. and T.E. McCurdy (1986): Instrumental Variable Estimation of an Error Components Model. *Econometrica,* 54, 869–881.

Anderson, T.W., and C. Hsiao (1981): Estimation of Dynamic Models with Error Components. *Journal of the American Statistical Association*, 76, 598–606.

Anderson, T.W., and C. Hsiao (1982): Formulation and Estimation of Dynamic Models Using Panel Data. *Journal of Econometrics*, 18, 47–82.

Arellano, M., and S. Bond (1991): Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, 58, 277–297.

Arellano, M., and O. Bover (1995): Another Look at the Instrumental Variable Estimation of Error-Components Models. *Journal of Econometrics*, 68, 29–51.

Avery, R.B. (1977): Error Component Models and Seemingly Unrelated Regressions. *Econometrica,* 45, 199–209.

Balestra, P. (1983): La Dérivation Matricielle. Collection de l'Institut de Mathématiques Economiques de Dijon, 12, Sirey, Paris.

Balestra, P. and J. (Varadharajan–)Krishnakumar (1987): Full Information Estimations of a System of Simultaneous Equations with Error Component Structure. *Econometric Theory,* 3, 223–246.

Baltagi, B.H. (1981): Simultaneous Equations with Error Components. *Journal of Econometrics,* 17, 189–200.

Baltagi, B.H. (1984): A Monte Carlo Study for Pooling Time Series of Cross-Section Data in the Simultaneous Equations Model. *International Economic Review*, 25, 603–624.

Baltagi, B.H. (2001): *Econometric Analysis of Panel Data*, second edition. Chichester: Wiley.

Baltagi, B.H. and Y-J. Chang (2000): Simultaneous Equations with Incomplete Panels, *Econometric Theory*, 16, 269–279.

Bekker, P., A. Kapteyn, and T. Wansbeek (1987): Consistent Sets of Estimates for Regressions with Correlated or Uncorrelated Measurement Errors in Arbitrary Subsets of All Variables. *Econometrica*, 55, 1223–1230.

Biørn, E. (1992): The Bias of Some Estimators for Panel Data Models with Measurement Errors. *Empirical Economics*, 17, 51–66.

Biørn, E. (1996): Panel Data with Measurement Errors. Chap. 10 in *The Econometrics of Panel Data. Handbook of the Theory with Applications*, ed. by L. Mátyás and P. Sevestre. Dordrecht: Kluwer.

Biørn, E. (2000): Panel Data with Measurement Errors. Instrumental Variables and GMM Estimators Combining Levels and Differences. *Econometric Reviews*, 19, 391–424.

Biørn, E. (2003): Handling the Measurement Error Problem by Means of Panel Data: Moment Methods Applied on Firm Data. Chap. 24 in *Econometrics and the Philosophy of Economics*, ed. by B. Stigum. Princeton: Princeton University Press.

Biørn, E., and T.J. Klette (1998): Panel Data with Errors-in-Variables: Essential and Redundant Orthogonality Conditions in GMM-Estimation. *Economics Letters*, 59, 275–282.

Biørn, E., and T.J. Klette (1999): The Labour Input Response to Permanent Changes in Output: An Errors in Variables Analysis Based on Panel Data. *Scandinavian Journal of Economics*, 101, 379–404.

Blundell, R., and S. Bond (1998): Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics*, 87, 115–143.

Bowden, R.J. and D.A. Turkington (1984): *Instrumental Variables*. Econometric Society Publication, No. 8, Cambridge University Press, Cambridge.

Breusch, T.S., G.E. Mizon and P. Schmidt (1987): Efficient Estimation Using Panel Data, Michigan State University Econometrics Workshop Paper 8608.
Breusch, T., H. Qian, P. Schmidt, and D. Wyhowski (1999): Redundancy of Moment Conditions. *Journal of Econometrics*, 91, 89–111.
Chamberlain, G.(1987): Asymptotic Efficiency in Estimation With Conditional Moment Restrictions. *Journal of Econometrics*, 34, 305–334.
Cornwell, C., P. Schmidt and D. Wyhowski (1992): Simultaneous Equations and Panel Data. *Journal of Econometrics*, 51, 151–181.
Davidson, R., and J.G. MacKinnon(1993): *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
Erickson, T. (1993): Restricting Regression Slopes in the Errors-in-Variables Model by Bounding the Error Correlation. *Econometrica*, 91, 959–969.
Fisher, F.M. (1966): *The Identification Problem in Econometrics*, New York: McGraw-Hill.
Fuller, W.A. (1987): *Measurement Error Models*. New York: Wiley.
Griliches, Z., and J.A. Hausman (1986): Errors in Variables in Panel Data. *Journal of Econometrics*, 31, 93–118.
Hansen, L.P. (1982): Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50, 1029–1054.
Harris, D. and L. Mátyás (1999): Introduction to the Generalized Method of Moments Estimation. Chap. 1 in *Generalized Method of Moments Estimation*, ed. by L. Mátyás. Cambridge: Cambridge University Press.
Harris, M., L. Mátyás and P. Sevestre, (2007): Dynamic Models for "Short Panels". Chap. 8 in this volume.
Hausman, J.A., and W.E. Taylor (1981): Panel Data and Unobservable Individual Effects. *Econometrica* 49, 1377–1398.
Hausman, J.A. and W.E. Taylor (1983): Identification in Linear Simultaneous Equations Models with Covariance Restrictions: An Instrumental Variables Interpretation. *Econometrica* 51, 1527–1549.
Holtz-Eakin, D., W. Newey, and H.S. Rosen (1988): Estimating Vector Autoregressions with Panel Data. *Econometrica*, 56, 1371–1395.
Hsiao, C. (2003): *Analysis of Panel Data*, 2nd edition. Cambridge: Cambridge University Press.
Klepper, S., and E. Leamer (1984): Consistent Sets of Estimates for Regressions with Errors in All Variables. *Econometrica*, 52, 163–183.
Koopmans, T.C. (1953): Identification Problems in Economic Model Construction, in *Studies in Econometric Method* (Cowles Commission Monograph 14), ed. by W.C. Hood and T.C. Koopmans, New York: John Wiley and Sons.
Krishnakumar, J. (1988): *Estimation of Simultaneous Equation Models with Error Components Structure*. Berlin, Heidelberg: Springer-Verlag.
Maddala, G.S. (1971): The Use of Variance Components Models in Pooling Cross Section and Time Series Data. *Econometrica,* 39, 341–358.
Mátyás, L. and L. Lovrics (1990): Small Sample Properties of Simultaneous Error Components Models. *Economics Letters* 32, 25–34.
McCabe, B., and A. Tremayne (1993): *Elements of Modern Asymptotic Theory with Statistical Applications*. Manchester: Manchester University Press.
Nagar, A.L. (1959): The Bias and Moment Matrix of the General k–class Estimators of the Parameters in Simultaneous Equations. *Econometrica,* 27, 575–595.
Nelson, C.R., and R. Startz (1990): Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator. *Econometrica*, 58, 967–976.
Nerlove, M. (1971): A Note on Error Components Models. *Econometrica,* 39, 383–396.
Newey, W.K. (1985): Generalized Method of Moments Specification Testing. *Journal of Econometrics*, 29, 229–256.
Newey, W.K., and D. McFadden (1994): Large Sample Estimation and Hypothesis Testing. Chap. 36 in *Handbook of Econometrics, Vol. IV*, ed. by R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.

Paterno, E.M., Y. Amemiya, and Y. Amemiya (1996): Random Effect and Random Coefficient Analysis with Errors-in-Variables. *1996 Proceedings of the Business and Economic Statistics Section*, pp. 76–79.

Phillips, P.C.B. (1982): Small Sample Distribution Theory in Econometric Models of Simultaneous Equations, Cowles Foundation Discussion Paper No. 617, Yale University.

Prucha, I.R. (1985): Maximum Likelihood and Instrumental Variable Estimation in Simultaneous Equation Systems with Error Components. *International Economic Review,* 26, 491–506.

Reiersøl, O. (1950): Identifiability of a Linear Relation Between Variables which are Subject to Error. *Econometrica*, 18, 375–389.

Rothenberg, T.J. (1971): Identification in Parametric Models. *Econometrica,* 39, 577–592.

Shalabh (2003): Consistent Estimation of Coefficients in Measurement Error Models with Replicated Observations. *Journal of Multivariate Analysis*, 86, 227–241.

Staiger, D., and J.H. Stock (1997): Instrumental Variables Regression With Weak Instruments. *Econometrica*, 65, 557–586.

Swamy, P.A.V.B. and S.S. Arora (1972): The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models. *Econometrica,* 40, 261–275.

Theil, H. (1971): *Principles of Econometrics*. Amsterdam: North-Holland Publishing Company.

Vella, F. and M. Verbeek (1999): Two-step Estimation of Panel Data Models with Censored Endogenous Variables and Selection Bias. *Journal of Econometrics*, 90, 239–263.

Wansbeek, T.J. (2001): GMM Estimation in Panel Data Models with Measurement Error. *Journal of Econometrics*, 104, 259–268.

Wansbeek, T.J., and R.H. Koning (1991): Measurement Error and Panel Data. *Statistica Neerlandica*, 45, 85–92.

Wansbeek, T.J., and E. Meijer (2000): *Measurement Error and Latent Variables in Econometrics*. Amsterdam: Elsevier.

White, H. (1982): Instrumental Variables Regression with Independent Observations. *Econometrica*, 50, 483–499.

White, H. (1984): *Asymptotic Theory for Econometricians*. Orlando: Academic Press.

White, H. (1986): Instrumental Variables Analogs of Generalized Least Squares Estimators. In *Advances in Statistical Analysis and Statistical Computing. Theory and Applications, vol. 1*, ed. by R.S. Mariano, Greenwich: JAI Press, pp. 173–227.

Zellner, A. and H. Theil (1962): Three Stage Least Squares: Simultaneous Estimation of Simultaneous Equations. *Econometrica,* 30, 54–78.

Ziliak, J.P. (1997): Efficient Estimation With Panel Data When Instruments Are Predetermined: An Empirical Comparison of Moment-Condition Estimators. *Journal of Business and Economic Statistics*, 15, 419–431.

Chapter 11
Pseudo-Panels and Repeated Cross-Sections

Marno Verbeek

11.1 Introduction

In many countries there is a lack of genuine panel data where specific individuals or firms are followed over time. However, repeated cross-sectional surveys may be available, where a random sample is taken from the population at consecutive points in time. Important examples of this are the Current Population Survey in the U.S.A., and the Family Expenditure Survey in the United Kingdom. While many types of models can be estimated on the basis of a series of independent cross-sections in a standard way, several models that seemingly require the availability of panel data can also be identified with repeated cross-sections under appropriate conditions. Most importantly, this concerns models with individual dynamics and models with fixed individual-specific effects.

Obviously, the major limitation of repeated cross-sectional data is that the same individuals are not followed over time, so that individual histories are not available for inclusion in a model, for constructing instruments or for transforming a model to first-differences or in deviations from individual means. All of these are often applied with genuine panel data. On the other hand, repeated cross-sections suffer much less from typical panel data problems like attrition and nonresponse, and are very often substantially larger, both in number of individuals or households and in the time period that they span.

In a seminal paper, Deaton (1985) suggests the use of cohorts to estimate a fixed effects model from repeated cross-sections. In his approach, individuals sharing some common characteristics (most notably year of birth) are grouped into cohorts, after which the averages within these cohorts are treated as observations in a pseudo panel. Moffitt (1993) and Collado (1997), in different ways, extend the approach of Deaton to nonlinear and dynamic models. Alternative estimators for

Marno Verbeek
Department of Financial Management, RSM Erasmus University, Burg. Oudlaan 50, 3062 PA Rotterdam, The Netherlands, e-mail: mverbeek@rsm.nl

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

the model with individual dynamics, including the one proposed by Girma (2000), are evaluated in Verbeek and Vella (2005). Alternative types of asymptotics are discussed in McKenzie (2004). In this chapter we shall discuss the identification and estimation of panel data models from repeated cross sections. In particular, attention will be paid to linear models with fixed individual effects, to models containing lagged dependent variables and to discrete choice models.

Models containing individual effects that are correlated with the explanatory variables ("fixed effects models") often arise naturally from economic theory, for example in life cycle models where the individual effects represent marginal utility of wealth (see, for example, Heckman and McCurdy (1980) or Browning, Deaton and Irish (1985)). Individual dynamics also often follow from economic theory, reflecting adjustment costs, habit persistence, or intertemporal optimization. Consequently, from an economic point of view it is important to be able to estimate dynamic models and models with fixed individual effects, even in the absence of genuine panel data. While it is possible to estimate such models using repeated cross-sections, we shall see below that such approaches typically require strong identification conditions, which are often hard to test.

Estimation techniques based on grouping individual data into cohorts are identical to instrumental variables approaches where the group indicators are used as instruments. Consequently, the grouping variables should satisfy the appropriate conditions for an instrumental variables estimator to be consistent (including a rank condition). This not only requires that the instruments are exogenous (in the sense of being uncorrelated to the unobservables in the equation of interest), but also relevant, i.e. appropriately correlated to the explanatory variables in the model. Loosely speaking, the latter requirement means that cohorts are defined as groups whose explanatory variables change differentially over time. Even if the instruments are exogenous and relevant, their large number and the fact that they may be only weakly correlated with the explanatory variables they are supposed to instrument may imply that the resulting estimators perform poorly because of the "weak instruments" problem (see Bound, Jaeger and Baker (1995), or Staiger and Stock (1997)).

The structure of this chapter is as follows. In Sect. 11.2 we present the basic linear model. Sect. 11.3 pays attention to linear dynamic models, while Sect. 11.4 briefly discusses the estimation of binary choice models. Sect. 11.5 concludes. A related survey can be found in Ridder and Moffitt (2007).

11.2 Estimation of a Linear Fixed Effects Model

We start with analyzing a simple linear model with individual effects given by

$$y_{it} = x_{it}'\beta + \alpha_i + u_{it},\ t = 1, \ldots, T, \tag{11.1}$$

where x_{it} denotes a K-dimensional vector of explanatory variables, and β is the parameter vector of interest. The index i refers to individuals and throughout this chapter we shall assume that the available data set is a series of independent

cross-sections, such that observations on N individuals are available in each period.[1] For simplicity, we shall assume that $E\{x_{it}u_{it}\} = 0$ for each t.

If the individual effects α_i are uncorrelated with the explanatory variables in x_{it}, the model in (11.1) can easily be estimated consistently from repeated cross-sections by pooling all observations and performing ordinary least squares treating $\alpha_i + u_{it}$ as composite error term. This exploits the K moment conditions in

$$E\{(y_{it} - x_{it}'\beta)x_{it}\} = 0. \tag{11.2}$$

However, in many applications the individual effects are likely to be correlated with some or all of the explanatory variables, so that at least some of the moment conditions in (11.2) are not valid. When genuine panel data are available, this can be solved using a fixed effects approach which treats α_i as fixed unknown parameters. In other words, each individual has its own intercept term. For estimating β, this is equivalent to using the within-transformed explanatory variables $x_{it} - \bar{x}_i$ as instruments for x_{it} in (11.1), where $\bar{x}_i = T^{-1}\sum_{t=1}^{T} x_{it}$. Obviously, when repeated observations on the same individuals are not available, such an approach cannot be used.

Deaton (1985) suggests the use of cohorts to obtain consistent estimators for β in (11.1) when repeated cross-sections are available, even if α_i is correlated with one or more of the explanatory variables. Let us define C cohorts, which are groups of individuals sharing some common characteristics. These groups are defined such that each individual is a member of exactly one cohort, which is the same for all periods. For example, a particular cohort may consist of all males born in the period 1950–1954. It is important to realize that the variables by which cohorts are defined should be observed for all individuals in the sample. This rules out time-varying variables (e.g. earnings), because these variables are observed at different points in time for the individuals in the sample. The seminal study of Browning, Deaton and Irish (1985) employs cohorts of households defined on the basis of five-year age bands subdivided as to whether the head-of-the-household is a manual or non-manual worker. Blundell, Duncan and Meghir (1998) employ year-of-birth intervals of 10 years, interacted with two education groups, Banks, Blundell and Preston (1994) use five-year age bands, while Propper, Rees and Green (2001) use 7 date of birth groups and 10 regions to construct cohorts.[2]

If we aggregate all observations to cohort level, the resulting model can be written as

$$\bar{y}_{ct} = \bar{x}_{ct}'\beta + \bar{\alpha}_{ct} + \bar{u}_{ct},\ c = 1,\ldots,C;\ t = 1,\ldots,T, \tag{11.3}$$

where $\bar{y}_{ct}$ is the average value of all observed y_{it}'s in cohort c in period t, and similarly for the other variables in the model. The resulting data set is a pseudo panel or synthetic panel with repeated observations over T periods and C cohorts. The main

[1] Because different individuals are observed in each period, this implies that i does not run from 1 to N for each t.

[2] Some authors employ the term "cohorts" to specifically reflect year-of-birth groups. We use "cohorts" in a broader sense, as groups of individuals (households, firms) sharing some common characteristics (most often including year-of-birth).

problem with estimating β from (11.3) is that $\bar{\alpha}_{ct}$ depends on t, is unobserved, and is likely to be correlated with $\bar{x}_{ct}$ (if α_i is correlated with x_{it}). Therefore, treating $\bar{\alpha}_{ct}$ as part of the random error term is likely to lead to inconsistent estimators. Alternatively, one can treat $\bar{\alpha}_{ct}$ as fixed unknown parameters assuming that variation over time can be ignored ($\bar{\alpha}_{ct} = \alpha_c$). If cohort averages are based on a large number of individual observations, this assumption seems reasonable and a natural estimator for β is the within estimator on the pseudo panel, given by

$$\hat{\beta}_W = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{x}_{ct} - \bar{x}_c)' \right)^{-1} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{y}_{ct} - \bar{y}_c), \tag{11.4}$$

where $\bar{x}_c = T^{-1} \sum_{t=1}^{T} \bar{x}_{ct}$ is the time average of the observed cohort means for cohort c. The properties of this estimator depend, among other things, upon the type of asymptotics that one is willing to employ. Deaton (1985) considers the asymptotic properties of this estimator when the number of cohorts C tends to infinity. This requires that the number of individuals N tends to infinity with (more or less) constant cohort sizes. Moffitt (1993), on the other hand, assumes that C is constant while the number of individuals tends to infinity. In this approach, cohort sizes tend to infinity, asymptotically.

The estimators proposed by Moffitt (1993) are based on the idea that grouping can be viewed as an instrumental variables procedure. To illustrate this, we shall reformulate the above estimator as an instrumental variables estimator based on a simple extension of (11.1). First, decompose each individual effect α_i into a cohort effect α_c and individual i's deviation from this effect. Letting $z_{ci} = 1$ ($c = 1, \ldots, C$) if individual i is a member of cohort c and 0 otherwise, we can write

$$\alpha_i = \sum_{c=1}^{C} \alpha_c z_{ci} + v_i, \tag{11.5}$$

which can be interpreted as an orthogonal projection. Defining $\alpha = (\alpha_1, \ldots, \alpha_C)'$ and $z_i = (z_{1i}, \ldots, z_{Ci})'$ and substituting (11.5) into (11.1), we obtain

$$y_{it} = x_{it}'\beta + z_i'\alpha + v_i + u_{it}. \tag{11.6}$$

If α_i and x_{it} are correlated, we may also expect that v_i and x_{it} are correlated. Consequently, estimating (11.6) by ordinary least squares would not result in consistent estimators. Now, suppose that instruments for x_{it} can be found that are uncorrelated with $v_i + u_{it}$. In this case, an instrumental variables estimator would typically produce a consistent estimator for β and α_c. A natural choice is to choose the cohort dummies in z_i, interacted with time, as instruments, in which case we derive linear predictors from the reduced forms

$$x_{k,it} = z_i'\delta_{kt} + w_{k,it}, \; k = 1, \ldots, K, \;\; t = 1, \ldots, T, \tag{11.7}$$

where δ_{kt} is a vector of unknown parameters. The linear predictor for x_{it} from this is given by $\hat{x}_{it} = \bar{x}_{ct}$, the vector of averages within cohort c in period t. The resulting instrumental variables estimator for β is then given by

$$\hat{\beta}_{\text{IV1}} = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c) x_{it}' \right)^{-1} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c) y_{it} \ , \tag{11.8}$$

which is identical to the standard within estimator based on the pseudo panel of cohort averages, given in (11.4).

The instrumental variables interpretation is useful because it illustrates that alternative estimators may be constructed using other sets of instruments. For example, if cohorts are constructed on the basis of age (year of birth), a more parsimonious function of age can be employed in (11.5) rather than a full set of age dummies. For example, z_i may include functions of year of birth, rather than a set of dummy variables. As argued by Moffitt (1993), it is likely that y_{it} will vary smoothly with cohort effects and, hence, those effects will be representable by fewer parameters than a full set of cohort dummies. Further, the instrument set in (11.7) can be extended to include additional variables. Most importantly however, the instrumental variables approach stresses that grouping data into cohorts requires grouping variables that should satisfy the typical requirements for instrument exogeneity and relevance. Basically, the approach of Deaton (1985) assumes that the cohort dummies, interacted with time dummies, provide valid instruments for all explanatory variables in the model (including the full set of cohort dummies). This requires that the instruments are uncorrelated with the equation's error term, and imposes a rank condition stating that the instruments are "sufficiently" correlated with each of the explanatory variables.

As mentioned above, the asymptotic behavior of pseudo panel data estimators can be derived using alternative asymptotic sequences. In addition to the two dimensions in genuine panel data (N and T), there are two additional dimensions: the number of cohorts C, and the number of observations per cohort n_c. We consider the following possibilities, which are typical for most studies:

1. $N \to \infty$, with C fixed, so that $n_c \to \infty$;
2. $N \to \infty$ and $C \to \infty$, with n_c fixed.
3. $T \to \infty$, with N, C fixed (so that n_c is also fixed);

McKenzie (2004) also considers asymptotic sequences where $T \to \infty$ and $n_c \to \infty$. Note that asymptotic theory is not meant as a guideline for how our estimators will behave when we get more data. Rather, we appeal to asymptotic theory when some dimension of the sample we already have is large enough for this to be appropriate. Whether or not asymptotic theory provides a reasonable approximation of the finite sample properties of pseudo panel data estimators is an empirical question, and many papers present Monte Carlo studies to obtain some insight into this issue.

The following list provides an overview of the sample sizes used in several important empirical papers.

	T	C	$\bar{n}_c$
Browning, Deaton and Irish (1985)	7	16	190
Banks, Blundell and Preston (1994)	20	11	354
Blundell, Browning and Meghir (1994)	17	9	520
Alessie, Devereux and Weber (1997)	14	5	>1000
Blundell, Duncan and Meghir (1998)	25	8	142
Propper, Rees and Green (2001)	19	70	80

For most applications either Type 1 or Type 2 asymptotics provides the most reasonable choice, and in many cases type 1 asymptotics is (implicitly or explicitly) employed. In the theoretical literature, Moffitt (1993) and Verbeek and Vella (2005) employ type 1 asymptotics, while Deaton (1985), Verbeek and Nijman (1993) and Collado (1997) employ Type 2 (with or without $T \to \infty$). Under Type 1 asymptotics, the fixed effects estimator based on the pseudo panel, $\hat{\beta}_W$, is consistent for β, provided that

$$\operatorname*{plim}_{n_c \to \infty} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{x}_{ct} - \bar{x}_c)' \tag{11.9}$$

is finite and invertible, and that

$$\operatorname*{plim}_{n_c \to \infty} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)\bar{\alpha}_{ct} = 0. \tag{11.10}$$

While the first of these two conditions is similar to a standard regularity condition, in this context it is somewhat less innocent. It states that the cohort averages exhibit genuine time variation, even with very large cohorts. Whether or not this condition is satisfied depends upon the way the cohorts are constructed, a point to which we shall return below.

Because $\bar{\alpha}_{ct} \to \alpha_c$, for some α_c if the number of observations per cohort tends to infinity, (11.10) will be satisfied automatically. Consequently, letting $n_c \to \infty$ and using Type 1 asymptotics is a convenient choice to arrive at a consistent estimator for β, see Moffitt (1993) and Ridder and Moffitt (2007). However, as argued by Verbeek and Nijman (1992) and Devereux (2007), even if cohort sizes are large, the small-sample bias in the within estimator on the pseudo panel may still be substantial.

Deaton (1985) proposes an alternative estimator for β that does not reply upon having a large number of observations per cohort, using Type 2 asymptotics.[3] A convenient starting point for this estimator is the cohort population version of (11.3), given by

$$y_{ct} = x_{ct}'\beta + \alpha_c + u_{ct},\ c = 1, \ldots, C;\ t = 1, \ldots, T, \tag{11.11}$$

[3] As argued by McKenzie (2004), in many applications cohorts are defined by age groups and hence a fixed number of cohorts is most likely to be of interest, which is inconsistent with Type 2 asymptotics. If $C \to \infty$ with $N \to \infty$, one needs to think of what this means for the distribution of population cohort means as well as the distribution of individual observations around these means. For example, it would be hard to argue that the covariance matrix on the right-hand side of (11.12) below is independent of how many cohorts are distinguished. See Verbeek and Nijman (1992) for more discussion and a Monte Carlo experiment that takes this issue into account.

where the variables denote unobservable population cohort means, and where α_c is the cohort fixed effect, which is constant because population cohorts contain the same individuals in each period. Now, $\bar{x}_{ct}$ and $\bar{y}_{ct}$ can be considered as error-ridden measurements of x_{ct} and y_{ct}. In particular, it is assumed that the measurement errors are distributed with zero mean, independent of the true values, i.e.

$$\begin{pmatrix} \bar{y}_{ct} - y_{ct} \\ \bar{x}_{ct} - x_{ct} \end{pmatrix} \sim IID\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} \sigma_{00} & \sigma' \\ \sigma & \Sigma \end{pmatrix} \right), \tag{11.12}$$

where the population cohort means are treated as fixed unknown constants. Although Σ, σ and σ_{00} are unknown, they can easily be estimated consistently (for N or T tending to infinity), using the individual data. Once estimates for Σ and σ are available, it is easy to adjust the moment matrices in the within estimator to eliminate the variance due to measurement error (cf. Fuller, 1987). This leads to the following errors-in-variables estimator

$$\begin{aligned} \hat{\beta}_D = & \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{x}_{ct} - \bar{x}_c)' - \tau \hat{\Sigma} \right)^{-1} \\ & \times \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{y}_{ct} - \bar{y}_c)' - \tau \hat{\sigma} \right), \end{aligned} \tag{11.13}$$

where $\hat{\Sigma}$ and $\hat{\sigma}$ are estimates of Σ and σ, respectively, and where $\tau = (T-1)/T$. As discussed in Verbeek and Nijman (1993), the original estimator presented by Deaton (1995) is characterized by $\tau = 1$. However, eliminating the incidental parameters (in α_c) first by within transforming the data, and working out the appropriate moments, suggests $\tau = (T-1)/T$, which leads to better small sample properties.

Under Type 1 asymptotics, the number of observations per cohort tends to infinity and both Σ and σ tend to zero, as well as their estimators. In this case $\hat{\beta}_D$ is asymptotically equivalent to $\hat{\beta}_W$. Accordingly, most empirical studies ignore the errors-in-variables problem and use standard estimators, like $\hat{\beta}_W$, see, for example, Browning, Deaton and Irish (1985), with an average cohort size of 190, or Blundell, Browning and Meghir (1994), with cohort sizes around 500. Unfortunately, there is no general rule to judge whether n_c is large enough to use asymptotics based on $n_c \to \infty$. Verbeek and Nijman (1992) analyze the bias in $\hat{\beta}_W$ for finite values of n_c. Depending upon the way in which the cohorts are constructed, the bias in the standard within estimator may still be substantial, even if cohort sizes are fairly large. In general, it holds that, for given n_c, the bias is smaller if the cohorts are chosen such that the relative magnitude of the measurement errors is smaller compared to the within cohort variance of x_{ct}. In practice, however, it may not be easy to construct cohorts in such a way. More recently, Devereux (2007) argues that cell sizes should be much larger, possibly 2000 or more.

In addition to the sizes of the cohorts, the way in which the cohorts are constructed is important. In general, one should be equally careful in choosing cohorts as in selecting instruments. In practice, cohorts should be defined on the basis of

variables that do not vary over time and that are observed for all individuals in the sample. This is a serious restriction. Possible choices include variables like age (date of birth), gender, race, or region.[4] Identification of the parameters in the model requires that the reduced forms in (11.7) generate sufficient variation over time. This requirement puts a heavy burden on the cohort identifying variables. In particular, it requires that groups are defined whose explanatory variables all have changed differentially over time.

Suppose, as an extreme example, that cohorts are defined on the basis of a variable that is independent of the variables in the model. In that case, the true population cohort means x_{ct} would be identical for each cohort c (and equal the overall population mean) and the only source of variation left in the data that is not attributable to measurement error would be the variation of x_{ct} over time. If these population means do not change over time, all variation in the observed cohort averages $\bar{x}_{ct}$ is measurement error and the errors-in-variables estimator $\hat{\beta}_D$ does not have a well-defined probability limit.

11.3 Estimation of a Linear Dynamic Model

An important situation where the availability of panel data seems essential to identify and estimate the model of interest is the case where a lagged dependent variable enters the model. Let us consider a simple extension of (11.1) given by

$$y_{it} = \gamma y_{i,t-1} + x_{it}'\beta + \alpha_i + u_{it},\ t = 1,\ldots,T, \tag{11.14}$$

where the K-dimensional vector x_{it} may include time-invariant and time-varying variables. When genuine panel data are available, the parameters γ and β can be estimated consistently (for fixed T and $N \to \infty$) using the instrumental variables estimators of Anderson and Hsiao (1981) or, more efficiently, using the GMM estimator of Arellano and Bond (1991). These estimators are based on first-differencing (11.14) and then using lagged values of $y_{i,t-1}$ as instruments.

In the present context, $y_{i,t-1}$ refers to the value of y at $t-1$ for an individual who is only observed in cross-section t. Thus, an observation for $y_{i,t-1}$ is unavailable. Therefore, the first step is to construct an estimate by using information on the y-values of other individuals observed at $t-1$. To do so, let z_i denote a set of time-invariant variables, including an intercept term. Now, consider the orthogonal projection in cross-section t of y_{it} upon z_i,

$$E^*\{y_{it}|z_i\} = z_i'\delta_{0t},\ t = 1,\ldots,T, \tag{11.15}$$

where E^* denotes the orthogonal projection (for a given t). This is similar to the reduced forms for $x_{k,it}$ in (11.7). Following Moffitt (1993), one obtains an estimate of $y_{i,t-1}$ as the predicted value from this regression, substituting the appropriate z

[4] Note that residential location may be endogenous in certain applications.

values for the individuals in cross-section t. That is,

$$\hat{y}_{i,t-1} = z_i' \hat{\delta}_{0,t-1}, \tag{11.16}$$

noting that $\hat{\delta}_{0,t-1}$ is estimated from data on different individuals than those indexed by i. In many circumstances it is convenient to think of z_i as a vector of dummy variables, corresponding to mutually exclusive cohorts, as in the previous section. In this case, the orthogonal projection in (11.15) corresponds to the conditional expectation and (11.16) corresponds to taking period-by-period sample averages within person i's cohort.

Now, insert these predicted values into the original model to get:

$$y_{it} = \gamma \hat{y}_{i,t-1} + x_{it}' \beta + \varepsilon_{i,t}, \; t = 1, \ldots, T; \tag{11.17}$$

where

$$\varepsilon_{it} = \alpha_i + u_{it} + \gamma (y_{i,t-1} - \hat{y}_{i,t-1}). \tag{11.18}$$

No matter how $\hat{y}_{i,t-1}$ is generated, its inclusion implies that one of the explanatory variables is measured with error, although the measurement error will be (asymptotically) uncorrelated with the predicted value.[5] To see whether it would be useful to estimate (11.17) by ordinary least squares, let us first of all make the assumption that the instruments in z_i are exogenous, so that

$$E\{(\alpha_i + u_{it}) z_i\} = 0, \; t = 1, \ldots, T. \tag{11.19}$$

This excludes the possibility that there are cohort effects in the unobservables. While this may appear unreasonable, this assumption is made in Moffitt (1993), Girma (2000) and in a number of cases in McKenzie (2004). Under (11.19) it can be argued that $\hat{y}_{i,t-1}$ and $\varepsilon_{i,t}$ are uncorrelated, which is a necessary condition for OLS applied to (11.17) to be consistent. In addition, consistency of OLS requires that x_{it} and ε_{it} are uncorrelated. This assumption may also be problematic, even in cases where the explanatory variables are exogenous to begin with, i.e. even if

$$E\{(\alpha_i + u_{it}) x_{it}\} = 0, \; t = 1, \ldots, T. \tag{11.20}$$

This is because x_{it} is likely to be correlated with $y_{i,t-1} - \hat{y}_{i,t-1}$.

Consider, for example, a case where high x-values in one period on average correspond with high x-values in the next period. If the β coefficients are positive this will generally imply that a high value for $x_{i,t-1}$, which is unobservable, will result in an underprediction of $y_{i,t-1}$. On the other hand, $x_{i,t-1}$ is positively correlated with x_{it}. Consequently, this will produce a positive correlation between ε_{it} and x_{it}, resulting in an inconsistent estimator for β. This inconsistency carries over to γ unless $\hat{y}_{i,t-1}$ is uncorrelated with x_{it}. As a result, the estimator suggested by Moffitt (1993), based on applying OLS to (11.17), is typically inconsistent unless there are either

[5] Unlike the standard textbook measurement error examples.

no time-varying exogenous regressors or the time-varying exogenous variables do not exhibit any serial correlation (see Verbeek and Vella, 2005).

To overcome the problem of correlation between the regressors and the error term in (11.17) one may employ an instrumental variables approach. Note that now we need instruments for x_{it} even though these variables are exogenous in the original model. Because these instruments will have to satisfy a condition like (11.19), a natural choice is to use the same instruments for x_{it} as we did for $y_{i,t-1}$. This will also guarantee that the instruments are uncorrelated with the prediction error $y_{i,t-1} - \hat{y}_{i,t-1}$ in ε_{it}.

As before, when the instruments z_i are a set of cohort dummies, estimation of (11.17) by instrumental variables is identical to applying OLS to the original model where all variables are replaced by their (time-specific) cohort sample averages. We can write this as

$$\bar{y}_{ct} = \gamma \bar{y}_{c,t-1} + \bar{x}'_{ct}\beta + \bar{\varepsilon}_{ct},\ c = 1, \ldots, C;\ t = 1, \ldots, T, \tag{11.21}$$

where all variables denote period-by-period averages within each cohort. For this approach to be appropriate, we need that $\bar{y}_{c,t-1}$ and $\bar{x}_{ct}$ are not collinear, which requires that the instruments capture variation in $y_{i,t-1}$ independent of the variation in x_{it}. That is, the time-invariant instruments in z_i should exhibit sufficient correlation with the exogenous variables in x_{it} and the (unobserved) lagged dependent variable $y_{i,t-1}$, while at the same time they should not be correlated with ε_{it}. Given these stringent requirements, it is likely that in many applications the number of available valid instruments is small. Verbeek and Vella (2005) provide more details on this rank condition.

The pairwise quasi-differencing approach of Girma (2000) deviates from the above estimation strategy in two respects, although it essentially makes the same assumptions. First, the lagged value of y is not approximated by the lagged cohort average but by an arbitrarily selected observation from the cohort. Second, the instruments are not the cohort dummies, but individual, or averaged, observations from the cohort. As a result, Girma's approach employs a noisy approximation to the unobserved lagged values as well as noisy instruments. Although, under appropriate assumptions, this noise will cancel out asymptotically, there does not seem to be any gain in using such an approach (see Verbeek and Vella (2005) for more discussion).

The availability of appropriate instruments satisfying condition (11.19) may be rather limited, because cohort effects in the unobservables are not allowed. It is possible to include cohort fixed effects in the model in essentially the same way as in the static case by including the cohort dummies z_i in the equation of interest, with time-invariant coefficients. This imposes (11.5) and results in

$$y_{it} = \gamma \hat{y}_{i,t-1} + x'_{it}\beta + z'_i\alpha + \eta_{it},\ t = 1, \ldots, T, \tag{11.22}$$

where

$$\eta_{it} = v_i + u_{it} + \gamma(y_{i,t-1} - \hat{y}_{i,t-1}), \tag{11.23}$$

and $E\{z_i v_i\} = 0$ by construction. This also allows us to relax (11.20) to

$$E\{(v_i + u_{it})x_{it}\} = 0,\ t = 1, \ldots, T. \tag{11.24}$$

Under these conditions, one would estimate (11.22) by instrumental variables using z_i, interacted with time dummies, as instruments. Verbeek and Vella (2005) refer to this as the augmented IV estimator noting that a time-varying α would make the model unidentified. To achieve identification, we need to assume that $\bar{y}_{c,t-1}$ and $\bar{x}_{ct}$ exhibit time variation and are not collinear. This condition puts additional restrictions upon the relationships between the instruments z_i and x_{it} and $y_{i,t-1}$. Among other things, at least three cross-sections are needed to identify the model under these assumptions.

Computation of this augmented IV estimator is remarkably simple if z_i is a set of cohort dummies. One simply aggregates the data into cohort averages, which gives

$$\bar{y}_{ct} = \gamma \bar{y}_{c,t-1} + \bar{x}'_{ct}\beta + \alpha_c + \bar{\eta}_{c,t}, \tag{11.25}$$

where $\alpha_c = z'_i\alpha$ denotes a cohort-specific fixed effect. Applying OLS to (11.25) corresponds to the standard within estimator for $(\gamma, \beta')'$ based upon treating the cohort-level data as a panel, which is consistent under the given assumptions (and some regularity conditions) under Type 1 asymptotics ($N \to \infty$ with C fixed). The usual problem with estimating dynamic panel data models (see Nickell (1981)),[6] does not arise because under assumption (11.24) the error term, which is a within cohort average of individual error terms that are uncorrelated with z_i, is asymptotically zero.[7] However, it remains to be seen whether suitable instruments can be found that satisfy the above conditions, because the rank condition for identification requires that the time-invariant instruments have time-varying relationships with the exogenous variables and the lagged dependent variable, while they should not have any time-varying relationship with the equation's error term. While this seems unlikely, it is not impossible. When z_i is uncorrelated with η_{it}, it is typically sufficient that the means of the exogenous variables, conditional upon z_i, are time-varying; see Verbeek and Vella (2005) for more details. Under Type 2 asymptotics ($N \to \infty$ with $C \to \infty$), we encounter similar problems as in the static case, and Collado (1997) discusses how this is handled in the dynamic model, by extending the approach of Deaton (1985). The resulting estimator is similar to the GMM-type estimators that are applied with genuine panel data (Arellano and Bond, 1991), but where the moment matrices are adjusted to reflect the errors-in-variables problem (for finite n_c).

Both Girma (2000) and McKenzie (2004) consider the linear dynamic model with cohort-specific coefficients in (11.14). While this extension will typically only make sense if there is a fairly small number of well-defined cohorts, it arises naturally from the existing literature on dynamic heterogeneous panels. For example,

[6] With genuine panel data, the within estimator in the dynamic model has a substantial bias for small and moderate values of T.

[7] Recall that, asymptotically, the number of cohorts is fixed and the number of individuals goes to infinity.

Robertson and Symons (1992) and Pesaran and Smith (1995) stress the importance of parameter heterogeneity in dynamic panel data models and analyze the potentially severe biases that may arise from handling it in an inappropriate manner. In many practical applications, investigating whether there are systematic differences between, for example, age cohorts, is an interesting question. Obviously, relaxing specification (11.14) by having cohort-specific coefficients puts an additional burden upon the identifying conditions. Further, note that using Type 2 asymptotics, where the number of cohorts increases with sample size, does not make much sense in these cases.

11.4 Estimation of a Binary Choice Model

In this section we briefly consider the estimation of a binary choice model on the basis of repeated cross-sections. In a binary choice model the outcome variable takes on only two different values, coded as 0 and 1. For example, the dependent variable could reflect whether or not a household owns a house, or whether or not an individual has a paid job. The model of interest is given by

$$y_{it}^* = x_{it}'\beta + \alpha_i + u_{it},\ t = 1,\ldots,T, \tag{11.26}$$

where y_{it}^* is a latent variable, and we observe

$$\begin{aligned} y_{it} &= 1 \text{ if } y_{it}^* > 0, \\ &= 0 \text{ otherwise.} \end{aligned} \tag{11.27}$$

With genuine panel data, popular parametric estimators for this model are the random effects probit estimator and the fixed effects logit estimator. The first approach assumes that the unobservables α_i and u_{it} are normally distributed and independent of the explanatory variables in x_{it}. The corresponding likelihood function takes into account that different observations on the same individual are dependent. With repeated cross-sections, this dependence is zero by construction and the binary choice probit model can be estimated as a pooled probit assuming $\alpha_i + u_{it}$ is $N(0,1)$.

Estimation becomes more complicated if one wants to allow α_i and x_{it} to be correlated, as in the fixed effects case. With genuine panel data, one option is to explicitly model this correlation, as in the Chamberlain (1984) approach, who proposes to parametrize the conditional expectation of α_i given the exogenous variables as a linear function of the x_{it}'s. That is,

$$E(\alpha_i | x_{i1},\ldots,x_{iT}) = x_{i1}'\lambda_1 + \ldots + x_{iT}'\lambda_T, \tag{11.28}$$

which allows us to write

$$\alpha_i = x_{i1}'\lambda_1 + \ldots + x_{iT}'\lambda_T + \xi_i, \tag{11.29}$$

where $E(\xi_i|x_{i1},\ldots,x_{iT}) = 0$. Substituting (11.29) into (11.26) produces

$$y_{it}^* = x_{i1}'\pi_{t1} + \ldots + x_{iT}'\pi_{tT} + \xi_i + u_{it},\ t = 1,\ldots,T, \tag{11.30}$$

where $\pi_{ts} = \beta + \lambda_s$ if $s = t$ and $\pi_{ts} = \lambda_s$ otherwise. Making distributional assumptions on ξ_i and u_{it} (e.g. normality) allows the application of standard maximum likelihood. However, when only repeated cross-sections are available, we do not observe the full history of the explanatory variables, as required in (11.29), and this approach is not feasible. Collado (1998) shows how this model can be estimated using cohort data, based on substituting the cohort specific means $\bar{x}_{c1},\ldots,\bar{x}_{cT}$ into (11.30). Using Type 2 asymptotics, with $C \to \infty$ and more or less fixed cohort sizes, this introduces an errors-in-variables problem in the equation. However, under normality the covariances between the explanatory variables and the disturbances are known functions of the variances of the measurement error (which can be identified from the individual data). Collado (1998) derives the corresponding probability that $y_{it} = 1$, which can be used to estimate π_t for each cross section t. Next, the structural parameters β (and λ) can be estimated using a minimum distance estimator. Note that y_{it}^* as well as y_{it} are not aggregated to cohort averages in this approach.

An alternative approach is proposed by Moffitt (1993) and is based on estimating the binary choice model by instrumental variables, where the cohort dummies (or other functions of the variables that define cohorts) are used as instruments. As before, this is based on Type 1 asymptotics (with C fixed and $N \to \infty$). Using (11.5), write the latent variable equation as

$$y_{it}^* = x_{it}'\beta + z_i'\alpha + v_i + u_{it},\ t = 1,\ldots,T. \tag{11.31}$$

Assuming, as before, that the cohort indicators, interacted with time, provide valid instruments, we can estimate the binary choice model by instrumental variables. This requires the assumption that $v_i + u_{it}$ is normally distributed; see Ridder and Moffitt (2007) for more details. Moffitt (1993) and Ridder and Moffitt (2007) also discuss extensions to discrete choice models with a lagged dependent variable.

11.5 Concluding Remarks

In this chapter we have briefly discussed the problem of estimating panel data models from a time series of independent cross-sections. In particular, attention was paid to the estimation of static fixed effects models, to dynamic models with individual effects and to binary choice models.

The approach proposed by Deaton (1985) is to divide the population into a number of cohorts, being groups of individuals sharing some common characteristics, and to treat the observed cohort means as error-ridden measurements of the population cohort means. The resulting estimator for the static linear model with fixed effects is a corrected within estimator based on the cohort aggregates. Moffitt (1993) extends the work of Deaton by considering a general instrumental

variables framework, of which explicit grouping is a special kind. While both approaches assume that $N \to \infty$ they differ in the assumptions about what happens to the cohorts when N increases. In Deaton's approach, the number of cohorts C increases with N (with more or less constant cohort sizes), while in Moffitt's approach, the number of cohorts (which is equivalent to the number of instruments) is fixed and cohort sizes increase with N. In this latter approach, the errors-in-variables problem disappears.

Both Moffitt (1993) and Collado (1997) consider the linear dynamic model, based on different types of asymptotics. As argued by Verbeek and Vella (2005), the fixed effects estimator based on the pseudo panel of cohort averages may provide an attractive choice, even when a lagged dependent variable is included in the model. This deviates from the genuine panel data case, where the standard fixed effects estimator suffers from a substantial small-T bias in dynamic models. A Monte Carlo experiment by Verbeek and Vella (2005) shows that the bias that is present in the within estimator for the dynamic model using genuine panel data (see Nickell (1981)), is much larger than what is found for similar estimators employed upon cohort aggregates.

However, an important issue in both the static and dynamic models is the validity and relevance of the instruments that are used to construct the cohorts. A necessary condition for consistency of most estimators is that all exogenous variables exhibit genuine time-varying cohort-specific variation. That is, the cohorts have exogenous variables that change differentially over time. While it is not obvious that this requirement will be satisfied in empirical applications, it is also not easy to check, because estimation error in the reduced form parameters may hide collinearity problems. That is, sample cohort averages may exhibit time-variation while the unobserved population cohort averages do not.

References

Alessie, R., M.P. Devereux and G. Weber (1997), Intertemporal Consumption, Durables and Liquidity Constraints: A Cohort Analysis, *European Economic Review*, 41, 37–59.

Anderson, T.W. and C. Hsiao (1981), Estimation of Dynamic Models with Error Components, *Journal of the American Statistical Association*, 76, 598–606.

Arellano, M. and S. Bond (1991), Some Test of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations, *Review of Economic Studies*, 58, 277–297.

Banks, J., R. Blundell and I. Preston (1994), Life-Cycle Expenditure Allocations and the Consumption Costs of Children, *European Economic Review,* 38, 1391–1410.

Blundell, R., M. Browning and C. Meghir (1994), Consumer Demand and the Life-Cycle Allocation of Household Expenditures, *Review of Economic Studies*, 61, 57–80.

Blundell, R., A. Duncan and C. Meghir (1998), Estimating Labor Supply Responses Using Tax Reforms, *Econometrica*, 66, 827–861.

Bound, J., D. Jaeger and R. Baker (1995), Problems with Instrumental Variables Estimation When the Correlation Between Instruments and the Endogenous Explanatory Variable is Weak, *Journal of the American Statistical Association*, 90, 443–450.

Browning, M., A. Deaton and M. Irish (1985), A Profitable Approach to Labor Supply and Commodity Demands over the Life Cycle, *Econometrica*, 53, 503–543.

Chamberlain, G. (1984), Panel Data, in: Z. Griliches and M.D. Intrilligator, eds., *Handbook of Econometrics, Volume 2*, Elsevier Science, North-Holland.
Collado, M.D. (1997), Estimating Dynamic Models from Time Series of Independent Cross-Sections, *Journal of Econometrics*, 82, 37–62.
Collado, M.D. (1998), Estimating Binary Choice Models from Cohort Data, *Investigaciones Económicas*, 22, 259–276.
Deaton, A. (1985), Panel Data from Time Series of Cross Sections, *Journal of Econometrics*, 30, 109–126.
Devereux, P. (2007), Small Sample Bias in Synthetic Cohort Models of Labor Supply, *Journal of Applied Econometrics*, 22, 839–848.
Fuller, W.A. (1987), *Measurement Error Models*, John Wiley and Sons, New York.
Girma, S. (2000), A Quasi-Differencing Approach to Dynamic Modelling from a Time Series of Independent Cross-Sections, *Journal of Econometrics*, 98, 365–383.
Heckman, J.J. and Th.E. McCurdy (1980), A Life-Cycle Model of Female Labour Supply, *Review of Economic Studies*, 47, 47–74.
McKenzie, D.J. (2004), Asymptotic Theory for Heterogeneous Dynamic Pseudo-Panels, *Journal of Econometrics,* 120, 235–262.
Moffitt, R. (1993), Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross-Sections, *Journal of Econometrics*, 59, 99–123.
Nickell, S. (1981), Biases in Dynamic Models with Fixed Effects, *Econometrica*, 49, 1417–1426.
Pesaran, M.H. and R. Smith (1995), Estimating Long-Run Relationships from Dynamic Heterogeneous Panels, *Journal of Econometrics*, 68, 79–113.
Propper, C., H. Rees and K. Green (2001), The Demand for Private Medical Insurance in the UK: A Cohort Analysis, *The Economic Journal*, 111, C180–C200.
Ridder, G and R. Moffitt (2007), *The Econometrics of Data Combination*, in: J.J. Heckman and E.E. Leamer, eds., *Handbook of Econometrics, Volume 6B*, Elsevier Science, North-Holland.
Robertson, D. and J. Symons (1992), Some Strange Properties of Panel Data Estimators, *Journal of Applied Econometrics*, 7, 175–189.
Staiger, D. and J.H. Stock (1997), Instrumental Variables Regressions with Weak Instruments, *Econometrica*, 65, 557–586.
Verbeek, M. and Th.E. Nijman (1992), Can Cohort Data Be Treated As Genuine Panel Data?, *Empirical Economics*, 17, 9–23.
Verbeek, M. and Th.E. Nijman (1993), Minimum MSE Estimation of a Regression Model with Fixed Effects from a Series of Cross-Sections, *Journal of Econometrics*, 59, 125–136.
Verbeek, M. and F. Vella (2005), Estimating Dynamic Models from Repeated Cross-Sections, *Journal of Econometrics*, 127, 83–102.

Chapter 12
Attrition, Selection Bias and Censored Regressions

Bo Honoré, Francis Vella and Marno Verbeek

12.1 Introduction

In micro-econometric applications issues related to attrition, censoring and non-random sample selection frequently arise. For example, it is quite common in empirical work that the variables of interest are partially observed or only observed when some other data requirement is satisfied. These forms of censoring and selectivity frequently cause problems in estimation and can lead to unreliable inference if they are ignored. Consider, for example, the problems which may arise if one is interested in estimating the parameters from a labor supply equation based on the examination of a panel data set and where one's objective is to make inferences for the whole population rather than only the sample of workers. The first difficulty that arises is that hours are generally only observed for individuals that work. In this way the hours measure is generally censored at zero and this causes difficulties for estimation as straightforward least squares methods, either over the entire sample or only the subsample of workers, are not generally applicable. Second, many of the explanatory variables of interest, such as wages, are also censored in that they are only observed for workers. Moreover, in these instances many of these variables may also be endogenous to labor supply and this may also create complications in estimation. While panel data are frequently seen as a way to overcome issues related to endogeneity as the availability of repeated observations on the same unit can allow the use of various data transformations to eliminate the cause of the endogeneity, in many instances the use of panel data can complicate matters. For example,

Bo Honoré
Department of Economics, Princeton University, Princeton, NJ 08544-1021, USA,
e-mail: honore@Princeton.EDU

Francis Vella
Department of Economics, Georgetown University, Washington DC, USA,
e-mail: fgv@georgetown.edu

Marno Verbeek
Department of Financial Management, RSM Erasmus University, Burg. Oudlaan 50, 3062 PA Rotterdam, The Netherlands, e-mail: mverbeek@rsm.nl

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

in many cases different observations are not independent because of the presence of individual or time-specific unobserved heterogeneity in the equation of interest, which needs to be accounted for. This increases computational demands, particularly in non-linear models. Also, when one has repeated observations on the same unit it may be the case that some units exit from the data. This last complication is known as attrition and this can further complicate estimation if this exit occurs in a non-random manner.

The aim of this chapter is to present an overview of panel data models involving sample selection, endogenous explanatory variables, censoring and attrition and discuss their estimation. We consider this chapter complementary to Verbeek and Nijman (1996) in that we do not provide a discussion of the various forms of non-response that can arise in panel data nor do we explore the issues related to ignorability of non-response. Readers interested in these issues should refer to Verbeek and Nijman (1996). Rather, in this chapter we choose to focus our attention on the more important estimators for empirical work and also review the more recent innovations in this literature. In doing so we focus not only on the standard parametric, likelihood based, procedures for models with sample selectivity and censoring but we also discuss the appropriate semi-parametric procedures which are available. In the following two sections we present the general model and provide a heuristic description of the issues related to selectivity and attrition. We then examine the applicability of standard linear regression based on random or fixed effects procedures for these models. Having established that these procedures are generally only applicable under very restrictive conditions, we focus on the various available alternative forms of estimation. Thus the subsequent sections deal with parametric and semi-parametric estimation of various forms of the model based on different distributional assumptions and different forms of censoring operating in the model. The chapter concludes with a short summary of some empirical applications which involve the use of panel data, sample selection, censoring or endogenous regressors.

12.2 Censoring, Sample Selection and Attrition

To discuss the models and estimators that we consider in this chapter we first present a general model. We then impose restrictions on this model to produce special cases of interest, and consider the various estimators which can be employed to estimate the parameters of interest for these special cases. The general structure of the models considered in this chapter is represented by the following system of simultaneous equations:

$$y_{it}^* = m_1(x_{it}, z_{it}, y_{i,t-1}; \theta_1) + u_{it} \,, \tag{12.1}$$

$$z_{it}^* = m_2(x_{it}, x_{1it}, z_{i,t-1}; \theta_2) + v_{it} \,, \tag{12.2}$$

$$z_{it} = h(z_{it}^*; \theta_3), \tag{12.3}$$

$$y_{it} = y_{it}^* \text{ if } g_t(z_{i1}, \ldots, z_{iT}) = 1 \,, \tag{12.4}$$

$$= 0 \text{ (or unobserved) otherwise} \,,$$

where i indexes individuals ($i = 1, \ldots, N$) and t indexes time ($t = 1, \ldots, T$); y_{it}^* and z_{it}^* are latent endogenous variables with observed counterparts y_{it} and z_{it}; x_{it} and x_{1it} are vectors of exogenous variables; m_1 and m_2 denote general functions characterized by the unknown parameters in θ_1 and θ_2, respectively. While we will generally focus on the case where we impose index restrictions on the conditional means, we write the model in the more general form by employing the unknown functions m_1 and m_2 to capture possible non-linearities. The mapping from the latent variable to its observed counterpart occurs through the censoring functions h and g_t noting that the former may depend on the unknown parameter vector θ_3. We will generally focus on the case where $h(.)$ is an indicator function producing the value 1 if $z_{it}^* > 0$, in which case there are no unknown parameters in the censoring process. However, when we consider the available two-step estimators we will also consider some popular alternative selection rules and these may involve the estimation of additional parameters.

The function g_t indicates that y_{it}^* may only be observed for certain values of $z_{i1}, \ldots, z_{iT}$. This includes sample selection where y_{it} is only observed if, for example, $z_{it} = 1$ or, alternatively in the balanced subsample case, if $z_{i1} = \ldots = z_{iT} = 1$. Alternatively, we will consider a special case of interest in which we replace the censoring mechanism in (12.4) with

$$y_{it} = y_{it}^* \cdot I(y_{it}^* > 0) \,, \tag{12.5}$$

where $I(.)$ is an indicator function operator which produces the value 1 if event (.) occurs and zero otherwise. The model which incorporates (12.4) as the censoring or selection rule corresponds with the sample selection model. The model with (12.5) as the censoring mechanism corresponds to the censored regression model.

The above model is very general and nests many models of interest as special cases. For example, it encompasses the static sample selection and censored regression models in which we only observe the dependent variable of primary interest for some subset of the data depending on the operation of a specific selection rule. The primary difference between these two, captured in the censoring processes, is that the sample selection model allows for different factors driving the censoring, z_{it}^*, and the variation in y_{it}^*. In this sense it is a double index model. In contrast, the censored regression model imposes that a single index explains the latent variable in the censoring decision and also the variation in y_{it}^*. The difference between the two is not only a statistical issue in that in many economic models for which panel data estimation is applicable it is possible that the selection rule is based on a different process than that generating variation in the primary dependent variable of interest.

The model also incorporates a potential role for dynamics in both the y equation and the censoring process. That is, while panel data are frequently seen as a mechanism for eliminating unobservables which create difficulties in estimation, an important feature and major attraction of panel data is that it provides the ability

to estimate the dynamics of various economic relationships based on individual behavior. Clearly this is generally not possible with cross-sectional data. Some of the estimators we discuss below provide some scope for estimating such relationships.

Note that an important feature of these models is related to identification. In many of the models that we consider it is possible to obtain identification of the parameters of interest by simply relying on non-linearities which arise from the distributional assumptions. In general, this is not an attractive, nor frequently accepted, means of identification. As these issues are frequently quite complicated we avoid such a discussion by assuming that the elements in the vector x_{1it} appear as explanatory variables in the selection equation (12.2) but are validly excluded from the primary equation (12.1). In this way the models are generally identified. Readers who are particularly interested in identification should examine the cited papers for a more detailed discussion.

A key aspect of any panel data model is the specification and treatment of its disturbances. We write the respective equations' errors as

$$u_{it} = \alpha_i + \varepsilon_{it} \tag{12.6}$$

$$v_{it} = \xi_i + \eta_{it} \tag{12.7}$$

which indicates that they comprise individual effects, α_i and ξ_i, and individual specific time effects, ε_{it} and η_{it}, which are assumed to be independent across individuals. This corresponds to the typical one-way error components model. Moreover, we allow the errors of the same dimension to be correlated across equations. In some instances we will assume that both the individual effects and the idiosyncratic disturbances can be treated as random variables, distributed independently of the explanatory variables. In such cases, we will often assume that the error components are drawn from known distributions. For many empirical applications, however, these assumption are not appropriate. For example, one may expect that some subset of the explanatory variables are potentially correlated with the one or both of the different forms of disturbances. Accordingly, it is common to treat the individual effects as fixed effects, which are potentially correlated with the independent variable, and we will consider the available procedures for estimating under such conditions. Second, while distributional assumptions are frequently useful from the sake of implementation, for many applications they may not be appropriate. As many of the procedures we examine are likelihood based any misspecification of the parametric component may lead to the resulting estimators being inconsistent. Thus, while we begin the analysis of each sub-model by making distributional assumptions regarding the disturbances we will also examine some semi-parametric estimators which do not rely on distributional assumptions. Finally, note that for the majority of models the parameters of primary interest are those contained in the vector θ_1, the variance σ_ε^2 and, when appropriate, σ_α^2. In some instances, however, there may be interest in the θ_2 vector.

12.3 Sample Selection and Attrition

Before focussing on the estimation of the parameters described in the above model it is useful to briefly discuss the problem of selection bias through (12.4), or the difficulties which arise from the problems with the presence of censoring via (12.5). It is also useful to consider the effect of sample attrition which captures the process by which some observations disappear, potentially non-randomly, from the sample over time. To illustrate these issues we will follow the discussion in Vella (1998) for the cross-sectional case which is based on the original motivation of Heckman (1974, 1979) and Gronau (1974). Assume that we are interested in examining the determinants of the wages of females, y_{it}^*, when we only observe the wages of females who are working a positive number of hours. To determine which factors influence wages we examine a panel data set of women where only a sub-sample are engaged in market employment and report wages for each period t. Moreover, assume that the sample comprises of three types; (i) those working and reporting wages for the entire period; (ii) those who fail to work for at least one period but remain in the sample for the entire period; and (iii) those who either work or do not work but do appear in the sample for the entire period. First, the differences between the workers and non-workers determines whether the issue of selection bias might arise. Second, the differences between those who remain in the sample or disappear from the sample determines whether the problem of attrition bias occurs. To illustrate how these biases may arise let us characterize each individual by both her endowments of observable and unobservable characteristics.

First assume that the working sub-sample is chosen randomly from the population. If the working sub-sample have similar endowments of characteristics as the non-working sample there is no reason to suspect selectivity bias will be induced by examining the working sample. That is, as the sample is randomly chosen the average characteristics, in terms of both observable and unobservables, of the working sample should be similar to the average characteristics of the population. The same is also true of attrition bias. That is, provided that individuals simply disappear from the sample in a random manner there is no reason to expect that the attrition affects the characteristics of the sample.

Let us now consider where the decisions to work or remain in the sample are no longer determined by random processes. In this instance, depending on the nature of these processes, the working and non-working samples that one observes potentially have different characteristics from each other and may have different characteristics from those who are no longer in the original sample. Sample selection bias arises when some component of the work decision is relevant to the wage determining process, while attrition bias results from the wage determining process not being independent of the decision to remain in the sample. That is, when some of the determinants of the work/attrition decisions are also influencing the wage. However, if the relationship between each of these respective decisions and the wage is purely

through the observables one can control for this by including the appropriate conditioning variables in the wage equation. That is, one is able to control for these potential biases by including the appropriate variables in the wage equation. Thus, sample selection or attrition bias will not arise purely on the basis of differences in observable characteristics.[1]

However, if we now assume the unobservable characteristics affecting the work/attrition decisions are correlated with the unobservable characteristics affecting the wage we generate a relationship between the work/attrition decisions and the process determining wages. Controlling for the observable characteristics when explaining wages is insufficient as some additional processes are influencing the wage. That is, the process determining whether an individual works and the process determining whether an individual remains in the sample are also affecting the wage. If these unobservable characteristics are correlated with the observables then the failure to include an estimate of the unobservables will lead to incorrect inference regarding the impact of the observables on wages. Thus a bias will be induced due to the sample selection and attrition.

This discussion highlights that sample selectivity bias operates through unobservable elements, and their correlation with observed variables. More explicitly, we can see that the presence of selection bias will be determined by the relationship between the two composite errors. It is driven by the correlations between u_{it} and v_{it} (or $v_{i1}, \ldots, v_{iT}$) and their components. Thus in the panel data case one can immediately observe that there are two ways in which the model can be contaminated with selectivity. First, there is the possibility that the individual effects are correlated across equations. Second, there is the possibility that the idiosyncratic disturbances are correlated across equations.

Attrition bias can also be seen as a special case of selection bias in that it arises by the same underlying process. Namely, the unobservable components of the decision to remain in the sample are correlated with the unobservable components of the work or wage equations. In this case, however, one has less flexibility in modelling the attrition as we will not observe the time varying determinants of attrition over the entire sample.

In general, sample selection and attrition problems may arise when a rule other than simple random sampling determines how sampling from the underlying population takes place. This selection rule may distort the representation of the true population and consequently distort inferences based on the observed data using standard methods. Distorting selection rules may be the outcome of self-selection of economic agents, non-response decisions of agents or the consequence of the sample survey process.

[1] This assumes that the inclusion of these observables is considered appropriate in a wage equation. For example, lagged wages may affect sample selection or attrition, while lagged wages are typically not included in a wage equation.

12.4 Sample Selection Bias and Robustness of Standard Estimators

One can easily illustrate the problems generated by the presence of attrition or selection bias by examining the properties of standard estimators for the primary equation where we estimate only over the sample of uncensored observations. To do so consider the simplest case of (12.1) where the dependent variable is written as a linear function of only the exogenous explanatory variables:

$$y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it} \ , \tag{12.8}$$

where we consider that each of the selection rules captured in (12.4) and (12.5) can be written as $z_{it} = 1$. In these instances one should consider that what corresponds to the OLS estimation in the cross-section is OLS estimation over the pooled sample for which y^*_{it} is observed. To illustrate the problems with such pooled estimation of (12.8) we can take expectations of (12.8) conditional upon y_{it} being observed, which gives

$$E(y_{it}|x_{it}, z_{it} = 1) = x'_{it}\beta + E(\alpha_i|x_{it}, z_{it} = 1) + E(\varepsilon_{it}|x_{it}, z_{it} = 1) \ , \tag{12.9}$$

noting that the last two terms will in general have non-zero values, which are potentially correlated with the x's, due to the dependence between α_i and ξ_i, and ε_{it} and ξ_{it}. These terms will, in general, be non-zero whenever $Pr\{z_{it} = 1|y_{it}, x_{it}\}$ is not independent of y_{it}. Accordingly, least squares estimation of (12.8) will lead to biased estimates of β due to this misspecification of the mean.

This above result is well known in the cross-sectional case and is a restatement of the results of Heckman (1979). In that paper the sample selection/censoring problem is shown to be related to the misspecification of the conditional mean. Heckman shows that to correct for this misspecification of the mean, an additional variable can be included, constructed through the use of parameters from an auxiliary equation, explaining the probability of censoring. However, given that in the panel data setting we have repeated observations on the individual one might think that the availability of panel data estimators which exploit the nature of the error structure might provide some scope to eliminate this bias without the use of such a variable. Accordingly, it is useful to discuss the properties of the standard fixed effects and random effects estimators in the linear model when the selection mechanism is endogenous. Thus we first consider estimation of (12.8) by the standard linear fixed effects or random effects procedures.

To consider these estimators we first introduce some additional notation. Observations on y_{it} are treated as available if $z_{it} = 1$ and missing if $z_{it} = 0$. We define $c_i = \prod_{t=1}^{T} z_{it}$, so that $c_i = 1$ if and only if y_{it} is observed for all t. The first estimators for β that we consider are the standard random effects estimators. Defining

$$\lambda_i = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T_i\sigma_\alpha^2}}$$

where $T_i = \sum_{t=1}^{T} z_{it}$ denotes the number of time periods y_{it} is observed, the random effects estimator based on the unbalanced panel (using all available cases) can be written as

$$\hat{\beta}_{RE}^{U} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} (x_{it} - \lambda_i \bar{x}_i)(x_{it} - \lambda_i \bar{x}_i)' \right)^{-1} \times \left(\sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} (x_{it} - \lambda_i \bar{x}_i)(y_{it} - \lambda_i \bar{y}_i) \right) \tag{12.10}$$

where $\bar{x}_i = T_i^{-1} \sum_{t=1}^{T} z_i x_{it}$ and $\bar{y}_i = T_i^{-1} \sum_{t=1}^{T} z_i y_{it}$ denote averages over the available observations. In some cases attention may be restricted to the balanced sub-panel comprising only those individuals that have completely observed records. The resulting random effects estimator is given by

$$\hat{\beta}_{RE}^{B} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} c_i (x_{it} - \lambda_i \bar{x}_i)(x_{it} - \lambda_i \bar{x}_i)' \right)^{-1} \times \left(\sum_{i=1}^{N} \sum_{t=1}^{T} c_i (x_{it} - \lambda_i \bar{x}_i)(y_{it} - \lambda_i \bar{y}_i) \right). \tag{12.11}$$

Note that all units for which $c_i = 1$ will have the same value for λ_i. Under appropriate regularity conditions, these two estimators are consistent for $N \to \infty$ if

$$E(\alpha_i + \varepsilon_{it} | z_i) = 0 \,, \tag{12.12}$$

where $z_i = (z_{i1}, \ldots, z_{iT})'$. This condition states that the two components of the error term in the model are mean independent of the sample selection indicators in z_i (conditional upon the exogenous variables). This appears to be a very strong condition and essentially implies that the selection process is independent of both of the unobservables in the model. One would suspect that for a large range of empirical cases this is unlikely to be true and this does not appear to be an attractive assumption to impose.

Given that the random effects estimator does not appear to be useful in the presence of selection bias it is worth focussing on the suitability of the fixed effects estimators of β. For the unbalanced panel the estimator can be written as

$$\hat{\beta}_{FE}^{U} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \times \left(\sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) \right), \tag{12.13}$$

while the corresponding estimator for the balanced sub-panel is given by

$$\hat{\beta}_{FE}^{B} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} c_i (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \times \left(\sum_{i=1}^{N} \sum_{t=1}^{T} c_i (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) \right). \tag{12.14}$$

Under appropriate regularity conditions, consistency of these two estimators requires that

$$E(\varepsilon_{it} - \bar{\varepsilon}_i | z_i) = 0\,, \tag{12.15}$$

where $\bar{\varepsilon}_i = T_i^{-1} \sum_{t=1}^{T} z_{it} \varepsilon_{it}$. Clearly, this indicates that the estimation over the subsample for which $z_{it} = 1$ will produce consistent estimates if the random component determining whether $z_{it} = 1$ is eliminated in the fixed effects transformation. That is, the unobservable component determining selection for each individual is time-invariant. While this may be true in certain instances it is likely that in many empirical examples such an assumption would not be reasonable as it imposes that the selection process is independent of the idiosyncratic errors.

This discussion illustrates that the conventional linear panel data estimators are inappropriate for the linear model with selection. The random effects estimator essentially requires that selection is determined outside the model while the fixed effects estimator imposes that, conditional on the individual effects, the selection process is determined outside the model. While the fixed effects estimator is more robust, it still is unsatisfactory for most empirical examples of panel data models with selectivity. Accordingly, we now begin to examine a range of estimators which handle the situation for which (12.12) and (12.15) are not satisfied.

12.5 Tobit and Censored Regression Models

The first model considered can be fully described by a subset of the equations capturing the general model outlined above. The model has the form

$$y_{it}^{*} = m_1(x_{it}, y_{i,t-1}; \theta_1) + u_{it}, \tag{12.16}$$

$$\begin{aligned} y_{it} &= y_{it}^{*} \text{ if } y_{it}^{*} > 0, \\ &= 0 \text{ (or unobserved) otherwise.} \end{aligned} \tag{12.17}$$

This considers a latent variable y_{it}^{*}, decomposed into a conditional mean depending upon x_{it} and possibly a lagged observed outcome $y_{i,t-1}$, and an unobserved mean zero error term u_{it}. The observed outcome equals the latent value if the latter is positive and zero otherwise. This model is the panel data extension of the tobit type I (under certain distributional assumptions) or censored regression model which is

commonly considered in cross-sectional analyses. A special case, which we do not consider here, arises when both y_{it} and x_{it} are unobserved if $y_{it}^* \leq 0$.

We now consider estimation of this standard censored regression model in (12.16) and (12.17) under different sets of assumptions. The simplest case arises when the lagged dependent variable is excluded from (12.16), and when ε_{it} is assumed to be drawn from a normal distribution, independent of the explanatory variables. We then consider the model where we allow for a lagged dependent variable. As we will see the estimation is somewhat more difficult because one has to incorporate the additional complications arising from the initial conditions. We then proceed to a consideration of the model where we relax the distributional assumptions that we impose on the error terms.

12.5.1 Random Effects Tobit

First, we consider the static tobit model, given by

$$y_{it}^* = m_1(x_{it}; \theta_1) + u_{it} \ ,$$

where the censoring rule is stated in (12.5)

$$\begin{aligned} &y_{it} = y_{it}^* \text{ if } y_{it}^* > 0, \\ &y_{it} = 0 \text{ if otherwise.} \end{aligned} \tag{12.18}$$

We also assume that u_{it} has mean zero and constant variance, independent of $(x_{i1}, \ldots, x_{iT})$. In order to estimate θ_1 by maximum likelihood we add an additional assumption regarding the joint distribution of $u_{i1}, \ldots, u_{iT}$. The likelihood contribution of individual i is the (joint) probability/density of observing the T outcomes $y_{i1}, \ldots, y_{iT}$, which is determined from the joint distribution of the latent variables $y_{i1}^*, \ldots, y_{iT}^*$ by integrating over the appropriate intervals. In general, this will imply T integrals, which in estimation are typically to be computed numerically. When $T = 4$ or more, this makes maximum likelihood estimation infeasible.

If the u_{it} are assumed to be independent, we have that

$$f(y_{i1}, \ldots, y_{iT} | x_{i1}, \ldots, x_{iT}; \vartheta_1) = \prod_t f(y_{it} | x_{it}; \vartheta_1) \ ,$$

where ϑ_1 contains all relevant parameters (including θ_1), which involves T one-dimensional integrals only (as in the cross-sectional case). This, however is highly restrictive. If, instead, we impose the error components assumption that $u_{it} = \alpha_i + \varepsilon_{it}$, where ε_{it} is i.i.d. over individuals and time, we can write the joint probability/density as

$$f(y_{i1},\ldots,y_{iT}|x_{i1},\ldots,x_{iT};\vartheta_1) = \int_{-\infty}^{\infty} f(y_{i1},\ldots,y_{iT}|x_{i1},\ldots,x_{iT},\alpha_i;\vartheta_1) f(\alpha_i)\mathrm{d}\alpha_i$$

$$= \int_{-\infty}^{\infty} \left[\prod_t f(y_{it}|x_{it},\alpha_i;\vartheta_1)\right] f(\alpha_i)\mathrm{d}\alpha_i, \qquad (12.19)$$

where f is generic notation for a density or probability mass function. This is a feasible specification that allows the error terms to be correlated across different periods, albeit in a restrictive way. The crucial step in (12.19) is that conditional upon α_i the errors from different periods are independent.

In principle arbitrary assumptions can be made about the distributions of α_i and ε_{it}. For example, one could assume that ε_{it} is i.i.d. normal while α_i has a Student t-distribution. However, this may lead to distributions for $\alpha_i + \varepsilon_{it}$ that are nonstandard and this is unattractive. Accordingly, it is more common to start from the joint distribution of $u_{i1},\ldots,u_{iT}$. We assume that the joint distribution of $u_{i1},\ldots,u_{iT}$ is normal with zero means and variances equal to $\sigma_\alpha^2 + \sigma_\varepsilon^2$ and $\mathrm{cov}\{u_{it},u_{is}\} = \sigma_\alpha^2$, $s \neq t$. This is the same as assuming that α_i is $\mathrm{NID}(0,\sigma_\alpha^2)$ and ε_{it} is $\mathrm{NID}(0,\sigma_\varepsilon^2)$. The likelihood function can then be written as in (12.19), where

$$f(\alpha_i) = \frac{1}{\sqrt{2\pi}\sigma_\alpha}\exp\left\{-\frac{1}{2}\left(\frac{\alpha_i}{\sigma_\alpha}\right)^2\right\}. \qquad (12.20)$$

and

$$f(y_{it}|x_{it},\alpha_i;\vartheta_1) = \frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\exp\left\{-\frac{1}{2}\left(\frac{y_{it}-m_1(x_{it};\theta_1)-\alpha_i}{\sigma_\varepsilon}\right)^2\right\} \text{ if } y_{it} > 0$$

$$= 1 - \Phi\left(\frac{m_1(x_{it};\theta_1)+\alpha_i}{\sigma_\varepsilon}\right) \text{ if } y_{it} = 0\,, \qquad (12.21)$$

where Φ denotes the standard normal cumulative density function. The latter two expressions are similar to the likelihood contributions in the cross-sectional case, with the exception of the inclusion of α_i in the conditional mean. The estimation of this model is identical to estimation of the tobit model in the cross-sectional setting except that we now have to account for the in clusion of the individual specific effect. As this individual effect is treated as a random variable, and the disturbances in the model are normally distributed, the above procedure is known as random effects tobit. Note that while we do not do so here, it would be possible to estimate many of the models considered in the survey of cross-sectional tobit models by Amemiya (1984) by allowing for an individual random effect.

12.5.2 Random Effects Tobit with Endogenous Explanatory Variables

The discussion of the random effects tobit model in the previous section assumed that the disturbances are independent of the explanatory variables. One useful extension of the model would be instances where some of the explanatory variables were treated as endogenous. This is similar to the cross-sectional model of Smith and Blundell (1986) who present a conditional ML estimator to account for the endogeneity of the explanatory variables.[2] The estimator simply requires estimating the residuals from the model for the endogenous explanatory and including them as an additional explanatory variable in the cross-sectional tobit likelihood function. Vella and Verbeek (1999) extend this to the panel case by exploiting the error components structure of the model. We now present this case where we assume the endogenous explanatory variable is fully observed. The model has the following form:

$$y_{it}^* = m_1(x_{it}, z_{it}; \theta_1) + \alpha_i + \varepsilon_{it} \tag{12.22}$$

$$z_{it} = m_2(x_{it}, x_{1it}, z_{i,t-1}; \theta_2) + \xi_i + \eta_{it} \tag{12.23}$$

$$y_{it} = y_{it}^* \cdot (y_{it}^* > 0) \tag{12.24}$$

The model's disturbances are assumed to be generated by the following distribution:

$$\begin{pmatrix} \alpha_i \iota + \varepsilon_i \\ \xi_i \iota + \eta_i \end{pmatrix} |_{X_i} \sim \text{NID}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 \iota\iota' + \sigma_\varepsilon^2 I & \sigma_{\alpha\xi} \iota\iota' + \sigma_{\varepsilon\eta} I \\ & \sigma_\xi^2 \iota\iota' + \sigma_\eta^2 I \end{pmatrix} \right) \tag{12.25}$$

where ι is a T-vector of ones. Exploiting this joint normality assumption allows us to write

$$E(u_{it}|X_i, v_i) = \tau_1 v_{it} + \tau_2 \bar{v}_i \,, \tag{12.26}$$

where $\tau_1 = \sigma_{\varepsilon\eta}/\sigma_\varepsilon^2$, $\tau_2 = T(\sigma_{\alpha\xi} - \sigma_{\varepsilon\eta}\sigma_\xi^2/\sigma_\varepsilon^2)/(\sigma_\eta^2 + T\sigma_\xi^2)$ and $\bar{v}_i = T^{-1}\sum_{t=1}^T v_{it}$.

As the endogenous explanatory variable is uncensored the conditional distribution of the error terms in (12.22) given z_i remains normal with an error components structure. Thus one can estimate the model in (12.22) and (12.24) conditional on the estimated parameters from (12.23) using the random effects likelihood function, after making appropriate adjustments for the mean and noting that the variances now reflect the conditional variances.

Write the joint density of $y_i = (y_{i1}, \ldots, y_{iT})'$ and z_i given X_i as:[3]

$$f(y_i|z_i, X_i; \vartheta_1, \vartheta_2) f(z_i|X_i; \vartheta_2) \,, \tag{12.27}$$

[2] Rivers and Vuong (1988) consider the cross-sectional probit counterpart of the Smith and Blundell (1986) procedure.

[3] When (12.23) is dynamic with an exogenous initial value z_{io}, (12.27) is valid if z_{io} is included in X_i. When the initial value is endogenous, we need to include z_{i0} in z_i.

where ϑ_1 denotes $(\theta_1, \sigma_\alpha^2, \sigma_\varepsilon^2, \sigma_{\alpha\xi}, \sigma_{\varepsilon\eta})$ and ϑ_2 denotes $(\theta_2, \sigma_\xi^2, \sigma_\eta^2)$. We first estimate ϑ_2 by maximizing the marginal likelihood function of the z_i's. Subsequently, the conditional likelihood function

$$\prod_i f(y_i|z_i, X_i; \vartheta_1, \hat{\vartheta}_2) \tag{12.28}$$

is maximized with respect to ϑ_1 where $\hat{\vartheta}_2$ denotes a consistent estimate of ϑ_2. The conditional distribution of y_i given z_i is multivariate normal with an error components structure. The conditional expectation can be derived directly from (12.26), substituting $v_{it} = z_{it} - m_2(x_{it}, x_{1it}, z_{i,t-1}; \theta_2)$, while the covariance structure corresponds to that of $v_{1i} + v_{2,it}$, where v_{1i} and $v_{2,it}$ are zero mean normal variables with zero covariance and variances

$$\sigma_1^2 = V\{v_{1i}\} = \sigma_\varepsilon^2 - \sigma_{\varepsilon\eta}^2 \sigma_\eta^{-2}\,, \tag{12.29}$$

$$\sigma_2^2 = V\{v_{2,it}\} = \sigma_\alpha^2 - \frac{T\sigma_{\alpha\xi}^2 \sigma_\eta^2 + 2\sigma_{\alpha\xi}\sigma_{\varepsilon\eta}\sigma_\eta^2 - \sigma_{\varepsilon\eta}^2 \sigma_\xi^2}{\sigma_\eta^2(\sigma_\eta^2 + T\sigma_\xi^2)}\,. \tag{12.30}$$

These follow from straightforward matrix manipulations and show that the error components structure is preserved and the conditional likelihood function of (12.22) and (12.24) has the same form as the marginal likelihood function without endogenous explanatory variables.[4]

The conditional maximum likelihood estimator can be extended to account for multiple endogenous variables as the appropriate conditional expectation is easily obtained as all endogenous regressors are continuously observed. Even if the reduced form errors of the endogenous regressors are correlated, provided they are characterized by an error components structure it can be shown that the conditional distribution of $\alpha_i + \varepsilon_{it}$ also has an error components structure. Time-specific heteroskedasticity in ε_{it} does not affect the conditional expectations and can be incorporated by having σ_1^2 vary over time. The model can also be estimated, along the lines suggested above, over subsets of the data chosen on the basis of z_{it}.

One obvious complication which arises in estimation of these models is that the standard errors have to be adjusted for the estimation uncertainty in the correction terms. This is an example of the standard "generated regressor problem" even though the second step is estimated by maximum likelihood. Vella and Verbeek (1999) provide the formulae for the standard errors in this particular context but for a more general treatment the reader is referred to Newey (1984).

In general the conditional maximum likelihood estimator cannot be employed when $z_{it} \neq z_{it}^*$. Thus the family of sample selection models considered below cannot be estimated by conditional maximum likelihood. One interesting exception,

[4] The algebraic manipulations are simplified if σ_1^2 and σ_2^2 replace the unconditional variances σ_ε^2 and σ_α^2 in ϑ_1. In this case, estimates for the latter two variances are easily obtained in a third step from the estimates from the first stage for σ_ξ^2 and σ_η^2, and the estimated covariances from the mean function, using the equalities in (12.29) and (12.30).

however, is when the primary equation is estimated over the subsample of individuals that have $z_{is} = z_{is}^*$, for all $s = 1, \ldots, T$.[5] This follows from the result that the error components structure is preserved when the reduced form dependent variables are observed.

Due to the presence of endogeneity in this model it is clear that one needs to carefully consider identification. In these models there is no non-linearity induced in the correction terms, but the non-linearity of m_1 or m_2 will identify the model. In the linear case, or if one does not want to rely on non-linearities for identification, exclusion restrictions are required. More explicitly, for each endogenous explanatory variable we need one exclusion restriction in the primary equation, unless, as before, the endogeneity can be restricted to be related to the time-invariant components only ($\sigma_{\varepsilon\eta} = 0$). This requires that x_{1it} is nonempty and has elements validly excluded from (12.22).

12.5.3 Dynamic Random Effects Tobit

The ability to estimate dynamic relationships from individual level data is an important attraction of panel data. Accordingly, an extension to the above model which involves the inclusion of a lagged dependent variable is of economic interest. Let us now reconsider the random effects tobit model, and generalize the latent variable specification to

$$y_{it}^* = m_1(x_{it}, y_{i,t-1}; \theta_1) + \alpha_i + \varepsilon_{it} \ , \tag{12.31}$$

with $y_{it} = y_{it}^*$ if $y_{it}^* > 0$ and 0 otherwise. Now consider maximum likelihood estimation of this dynamic random effects tobit model, making the same distributional assumptions as above. In general terms, the likelihood contribution of individual i is given by (compare (12.19))

$$\begin{aligned} f(y_{i1}, \ldots, y_{iT} | x_{i1}, \ldots, x_{iT}; \vartheta_1) &= \int_{-\infty}^{\infty} f(y_{i1}, \ldots, y_{iT} | x_{i1}, \ldots, x_{iT}, \alpha_i; \vartheta_1) f(\alpha_i) \mathrm{d}\alpha_i \\ &= \int_{-\infty}^{\infty} \left[\prod_{t=2}^{T} f(y_{it} | y_{i,t-1}, x_{it}, \alpha_i; \vartheta_1) \right] f(y_{i1} | x_{i1}, \alpha_i; \vartheta_1) f(\alpha_i) \mathrm{d}\alpha_i \ , \end{aligned} \tag{12.32}$$

where

$$\begin{aligned} &f(y_{it} | y_{i,t-1}, x_{it}, \alpha_i; \vartheta_1) \\ &= \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp\left\{ -\frac{1}{2} \left(\frac{y_{it} - m_1(x_{it}, y_{i,t-1}; \theta_1) - \alpha_i}{\sigma_\eta} \right)^2 \right\} \text{ if } y_{it} > 0, \\ &= 1 - \Phi\left(\frac{m_1(x_{it}, y_{i,t-1}; \theta_1) + \alpha_i}{\sigma_\varepsilon} \right) \text{ if } y_{it} = 0. \end{aligned}$$

[5] A similar argument is exploited in Arellano et al. (1999).

This is completely analogous to the static case and $y_{i,t-1}$ is simply included as an additional explanatory variable. However, the term $f(y_{i1}|x_{i1},\alpha_i;\theta_1)$ in the likelihood function may cause problems. It gives the distribution of y_{i1} without knowing its previous value but conditional upon the unobserved heterogeneity term α_i.

If the initial value is exogenous in the sense that its distribution does not depend upon α_i, we can place the term $f(y_{i1}|x_{i1},\alpha_i;\vartheta_1) = f(y_{i1}|x_{i1};\vartheta_1)$ outside the integral. In this case, we can simply consider the likelihood function conditional upon y_{i1} and ignore the term $f(y_{i1}|x_{i1};\vartheta_1)$ in estimation. The only consequence may be a loss of efficiency if $f(y_{i1}|x_{i1};\vartheta_1)$ provides information about ϑ_1. This approach would be appropriate if the starting value is necessarily the same for all individuals or if it is randomly assigned to individuals.

However, it may be hard to argue in many applications that the initial value y_{i1} is exogenous and does not depend upon a person's unobserved heterogeneity. In that case we would need an expression for $f(y_{i1}|x_{i1},\alpha_i;\vartheta_1)$ and this is problematic. If the process that we are estimating has been going on for a number of periods before the current sample period, $f(y_{i1}|x_{i1},\alpha_i;\vartheta_1)$ is a complicated function that depends upon person i's unobserved history. This means that it is typically impossible to derive an expression for the marginal probability $f(y_{i1}|x_{i1},\alpha_i;\vartheta_1)$ that is consistent with the rest of the model. Heckman (1981) suggests an approximate solution to this initial conditions problem that seems to work reasonably well in practice. It requires an approximation for the marginal distribution of the initial value by a tobit function using as much pre-sample information as available, without imposing restrictions between its coefficients and the structural parameters in θ_1. Vella and Verbeek (1998, 1999) provide illustrations of this approach. Wooldridge (2005) suggests an alternative approach that is based on considering the likelihood function conditional upon the initial values y_{i1} and then making parametric assumptions about the distribution of the unobserved effect conditional upon the initial value and any exogenous explanatory variables, $f(\alpha_i|y_{i1},x_{i1};\vartheta_1)$, rather than $f(y_{i1}|x_{i1},\alpha_i;\vartheta_1)f(\alpha_i)$. Because the impact of the initial conditions diminishes if the number of sample periods T increases, one may decide to ignore the problem when T is fairly large.

12.5.4 Fixed Effects Tobit Estimation

The fully parametric estimation of the tobit model assumes that both error components have a normal distribution, independent of the explanatory variables. Clearly, this is restrictive and a first relaxation arises if we treat the individual-specific effects α_i as parameters to be estimated, as is done in the linear fixed effects model. However, such an approach is generally not feasible in non-linear models. The log-likelihood function for the fixed effects tobit model has the general form

$$\log L = \sum_{i=1}^{N} \left[\sum_{t=1}^{T} \log f(y_{it}|x_{it},\alpha_i;\vartheta_1) \right], \tag{12.33}$$

where (as before)

$$f(y_{it}|x_{it},\alpha_i;\vartheta_1) = \frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\exp\left\{-\frac{1}{2}\left(\frac{y_{it}-m_1(x_{it};\theta_1)-\alpha_i}{\sigma_\varepsilon}\right)^2\right\} \text{ if } y_{it}>0$$
$$= 1-\Phi\left(\frac{m_1(x_{it};\theta_1)+\alpha_i}{\sigma_\varepsilon}\right) \text{ if } y_{it}=0.$$

Maximization of (12.33) can proceed through the inclusion of N dummy variables to capture the fixed effects or using an alternative strategy, described in Greene (2004), which bypasses the large computation demands of including so many additional variables.

This fixed effects tobit estimators is subject to the incidental parameter problem (Neyman and Scott, 1948, Lancaster, 2000), and results in inconsistent estimators of the parameters of interest if the number of individuals goes to infinity while the number of time periods is fixed. It was generally believed that the bias resulting from fixed effects tobit was large although more recent evidence provided by Greene suggests this may not be the case. On the basis of Monte Carlo evidence, Greene (2004) concludes that there is essentially no bias in the estimates of θ_1. However, the estimate of σ_ε is biased and this generates bias in the estimates of the marginal effects. Greene also concludes that the bias is small if T is 5 or greater.

Hahn and Newey (2004) suggest two approaches to bias reduction in fixed effects estimation of non-linear models such as the fixed effects tobit model. The first procedure is based on the use of jackknife methods and exploits the variation in the fixed effects estimator when each of the observations are, in turn, separately deleted. By doing so one is able to form a bias-corrected estimator using the Quenouille (1956) and Tukey (1958) jackknife formula. For simplicity, let $m_1(x_{it};\theta_1) = x_{it}'\beta$ and let $\widehat{\beta}_{(t)}$ denote the fixed effects estimator based on the subsample excluding the observations for the tth wave. The jackknife estimator ($\widehat{\beta}_{JK}$) is defined to be

$$\widehat{\beta}_{JK} = T\widehat{\beta} - (T-1)\sum_{t=1}^{T}\widehat{\beta}_{(t)}/T\ ,$$

where $\widehat{\beta}$ is the fixed effects estimator based on the entire panel. Hahn and Newey note that the panel jackknife is not particularly complicated. While it does require $(T+1)$ fixed effects estimations of the model one can employ the algorithm proposed by Greene, discussed above, and the estimates of $\widehat{\beta}$ and $\widehat{\alpha}_i$ can be used as starting values.

The second procedure is an analytic bias correction using the bias formula obtained from an asymptotic expansion as the number of periods grows. This is based on an approach suggested by Waterman et al. (2000) and is also related to the approach adopted by Woutersen (2002). Note that while none of these authors examine the fixed effect tobit model, preferring to focus mainly on discrete choice models, the approaches are applicable. Hahn and Newey (2004) provide some simulation evidence supporting the use of their procedures in the fixed effects probit model.

12.5.5 Semi-parametric Estimation

As shown in Honoré (1992) is also possible to estimate the parameters of panel data tobit models like (12.16) and (12.17) with no assumptions on the distribution of the individual specific effects and with much weaker assumptions on the transitory errors.

To fix ideas, consider a model with a linear index restriction, that is

$$y_{it}^{*} = x_{it}'\beta + \alpha_i + \varepsilon_{it} \ ,$$

and

$$\begin{aligned} y_{it} &= y_{it}^{*} \text{ if } y_{it}^{*} > 0, \\ y_{it} &= 0 \text{ otherwise.} \end{aligned}$$

The method proposed in Honoré (1992) is based on a comparison of any two time periods, t and s. The key insight behind the estimation strategy is that if ε_{it} and ε_{is} are identically distributed conditional on (x_{it}, x_{is}) then

$$\begin{aligned} v_{ist}(\beta) &= \max\{y_{is}, (x_{is} - x_{it})'\beta\} - \max\{0, (x_{is} - x_{it})'\beta\} \\ &= \max\{\alpha_i + \varepsilon_{is}, -x_{is}'\beta, -x_{it}'\beta\} - \max\{-x_{is}'\beta, -x_{it}'\beta\} \end{aligned}$$

and

$$\begin{aligned} v_{its}(\beta) &= \max\{y_{it}, (x_{it} - x_{is})'\beta\} - \max\{0, (x_{it} - x_{is})'\beta\} \\ &= \max\{\alpha_i + \varepsilon_{it}, -x_{it}'\beta, -x_{is}'\beta\} - \max\{-x_{it}'\beta, -x_{is}'\beta\} \end{aligned}$$

are also identically distributed conditional on (x_{it}, x_{is}). This can be used to construct numerous moment conditions of the form

$$E\left[\left(g(v_{ist}(\beta)) - g(v_{its}(\beta))\right)h(x_{it}, x_{is})\right] = 0 \tag{12.34}$$

If g is increasing and $h(x_{it}, x_{is}) = x_{is} - x_{it}$, these moment conditions can be turned into a minimization problem which identifies β subject to weak regularity conditions. For example, with $g(d) = d$, (12.34) corresponds to the first-order conditions of the minimization problem

$$\begin{aligned} \underset{b}{\text{minimize}}\ E\big[&\big(\max\{y_{is}, (x_{is} - x_{it})'b\} \\ &- \max\{y_{it}, -(x_{is} - x_{it})'b\} - (x_{is} - x_{it})'b\big)^2 \\ &+ 2 \cdot 1\{y_{is} < (x_{is} - x_{it})'b\}((x_{is} - x_{it})'b - y_{is})y_{it} \\ &+ 2 \cdot 1\{y_{it} < -(x_{is} - x_{it})'b\}(-(x_{is} - x_{it})'b - y_{it})y_{is}\big] \end{aligned}$$

which suggests estimating β by minimizing

$$\sum_{i=1}^{n}\sum_{s<t}\big(\max\{y_{is},(x_{is}-x_{it})'b\}$$
$$-\max\{y_{it},-(x_{is}-x_{it})'b\}-(x_{is}-x_{it})'b\big)^2$$
$$+2\cdot 1\{y_{is}<(x_{is}-x_{it})'b\}((x_{is}-x_{it})'b-y_{is})y_{it}$$
$$+2\cdot 1\{y_{it}<-(x_{is}-x_{it})'b\}(-(x_{is}-x_{it})'b-y_{it})y_{is} \qquad (12.35)$$

The objective function in (12.35) is convex in b, as are other objective functions based on (12.34). This means that it is extremely easy to find the estimator $\widehat{\beta}$. [Charlier et al. (2000) consider a conditional moment conditions estimator based on (12.34). This estimator is more efficient that then one obtained by minimizing (12.35) but it is more difficult to calculate.

Honoré and Kyriazidou (2000) discuss estimators defined by a general $g(d)$ as well as estimators based on moment conditions that are derived under the stronger assumption that the distribution of $(\varepsilon_{it},\varepsilon_{is})$ is exchangeable conditional on (x_{it},x_{is}).

12.5.6 Semi-parametric Estimation in the Presence of Lagged Dependent Variables

Honoré (1993), Hu (2002) and Honoré and Hu (2004) show how one can modify the moment conditions in (12.34) in such a way that one can allow for lagged dependent variables as explanatory variables. The specifics for this differs depending on whether the lagged latent or the lagged censored variable is used, and the main difficulty in this literature is that it is not easy to show that the moment conditions are uniquely satisfied at the true parameter values.

12.6 Models of Sample Selection and Attrition

As discussed above the tobit model has the somewhat unattractive feature that the index that explains the censoring also is required to explain the variation in the dependent variable of primary interest. We now turn our attention to the estimation of the model where the selection process is driven by a different index to that generating the dependent variable of primary interest. One might think, for example, that the number of hours an individual works depends on some different factors than those determining the work decision. At the very least one might think that the weights on each of the factors differs across decisions. For the sake of simplification we introduce the following form of the model

$$y_{it}^* = x_{it}'\beta + \alpha_i + \varepsilon_{it}, \tag{12.36}$$
$$z_{it}^* = x_{it}'\theta_{21} + x_{1it}'\theta_{22} + \xi_i + v_{it}, \tag{12.37}$$
$$z_{it} = I(z_{it}^* > 0), \tag{12.38}$$
$$y_{it} = y_{it}^* \cdot z_{it}$$

where we again highlight that the vector x_{1it} is nonempty (and not collinear with x_{it}). While this is sometimes seen as a controversial assumption we do not discuss the merits of such an approach. Below, we also discuss the scope of introducing dynamics into the primary equation.

12.6.1 Maximum Likelihood Estimators

Given that we can make distributional assumptions regarding the error components it is natural to construct a maximum likelihood estimator for all the parameters in (12.36) and (12.38). Consider the case where the individual effect is treated as a random effect and the disturbances are all normally distributed. To derive the likelihood function of the vectors z_i and y_i, we first write

$$\log f(z_i, y_i) = \log f(z_i|y_i) + \log f(y_i) \tag{12.39}$$

where $f(z_i|y_i)$ is the likelihood function of a conditional T-variate probit model and $f(y_i)$ is the likelihood function of a T_i-dimensional error components regression model, where $T_i = \sum_t z_{it}$. The second term can be written as

$$\begin{aligned}\log f(y_i) = &\frac{-T_i}{2}\log 2\pi - \frac{T_i - 1}{2}\log\sigma_\varepsilon^2 - \frac{1}{2}(\sigma_\varepsilon^2 + T_i\sigma_\alpha^2) \\ &- \frac{1}{2\sigma_\varepsilon^2}\sum_{t=1}^{T} z_{it}(y_{it} - x_{it}'\beta)^2 - \frac{T_i}{2(\sigma_\varepsilon^2 + T_i\sigma_\alpha^2)}(\bar{y}_i - \bar{x}_i'\beta)^2.\end{aligned} \tag{12.40}$$

The first term in (12.39) requires the derivation of the conditional distribution of the error term in the probit model. From the assumption of joint normality and defining $\pi_{it} = z_{it}(\alpha_i + \varepsilon_{it})$, the conditional expectation of $v_{it} = \xi_i + \eta_{it}$ is given by

$$\begin{aligned}E(\xi_i + \eta_{it}|\pi_{i1}, \ldots, \pi_{iT}) = &z_{it}\frac{\sigma_{\varepsilon\eta}}{\sigma_\varepsilon^2}\left[\pi_{it} - \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2 + T_i\sigma_\alpha^2}\sum_{t=1}^{T}\pi_{it}\right] \\ &+ \frac{\sigma_{\alpha\xi}}{\sigma_\varepsilon^2 + T_i\sigma_\alpha^2}\sum_{t=1}^{T}\pi_{it}\end{aligned} \tag{12.41}$$

Using our distributional assumptions the conditional distribution of $\xi_i + \eta_{it}$ given $\pi_{i1}, \ldots, \pi_{iT}$ corresponds to the unconditional distribution of the sum of three normal variables $e_{it} + \omega_{1i} + z_{it}\omega_{2i}$ whose distribution is characterized by

$$
\begin{aligned}
E(\omega_{1i}) = E(\omega_{2i}) = 0, E(e_{it}) &= c_{it} \\
V(e_{it}) &= \sigma_\eta^2 - z_{it}\sigma_{\varepsilon\eta}^2/\sigma_\varepsilon^2 = s_t^2 \\
V(\omega_{1i}) &= \sigma_\xi^2 - T_i\sigma_{\alpha\xi}^2(\sigma_\varepsilon^2 + T_i\sigma_\alpha^2)^{-1} = k_1 \\
V(\omega_{2i}) &= \sigma_{\varepsilon\eta}^2\sigma_\alpha^2\sigma_\varepsilon^{-2}(\sigma_\varepsilon^2 + T_i\sigma_\alpha^2)^{-1} = k_2 \\
\mathrm{cov}(\omega_{1i}, \omega_{2i}) &= -\sigma_{\alpha\xi}\sigma_{\varepsilon\eta}(\sigma_\varepsilon^2 + T_i\sigma_\alpha^2)^{-1} = k_{12},
\end{aligned}
$$

where the other covariances are all zero and note that we do not explicitly add an index i to the variances s_t^2, k_1 and k_2. Similar to the unconditional error components probit model the likelihood contribution can be written as

$$
\begin{aligned}
f(z_i|y_i) = \int\int \prod_{t=1}^{T} \Phi\left(d_{it} \frac{x_{it}'\theta_{21} + x_{1it}'\theta_{22} + c_{it} + \omega_{1i} + z_{it}\omega_{2i}}{s_t} \right) \\
\times f(\omega_{1i}, \omega_{2i}) d\omega_{1i} d\omega_{2i}
\end{aligned}
$$

where $d_{it} = 2z_{it} - 1$ and $f(.,.)$ is the density of ω_{1i} and ω_{2i}. Using these various expressions it is now possible to construct the complete likelihood function. Computation of the maximum likelihood estimator requires numerical integration over two dimensions for all individuals which are not observed in each period. Thus the computational demands are reasonably high and as a result this approach has not been proven to be popular in empirical work.

12.6.2 Two-Step Estimators

A shortcoming with the maximum likelihood approach outlined above is that the model can be sometimes difficult to estimate due to its computational demands. While there is relatively little experience of estimating such models in the panel setting (see Keane et al., 1988, for an example) it is clear that the ML selection type corrections in the cross-section setting is far less popular than the subsequently developed two-step estimators. To present the two-step estimators in the panel setting we follow the approach of Vella and Verbeek (1999). In this case we again start with the model presented in (12.36), (12.37) and (12.38). Note that although we focus on estimating the above model we retain some degree of generality. This allows us to more easily talk about extensions of the above model to alternative forms of censoring. The approach that we adopt is a generalization of the Heckman (1979) cross-sectional estimator to the panel data model. For the model that immediately follows, the estimation procedure is also found in Ridder (1990) and Nijman and Verbeek (1992).

To motivate a two-step estimator in this setting we begin by conditioning (12.36) on the vector z_i (and the matrix of exogenous variables X_i) to get

$$
E(y_{it}|X_i, z_{i0}, z_i) = x_{it}'\beta + E(u_{it}|X_i, z_{i0}, z_i)\,. \tag{12.42}
$$

If the mean function of (12.37) does not depend upon $z_{i,t-1}$ and sample selection only depends on the current value of z_{it}, it is possible to condition only on z_{it} and not $z_i = (z_{i1} \ldots . z_{iT})$ and base estimators on the corresponding conditional moments (see Wooldridge, 1995). In this case z_{i0} drops from the conditioning set. We assume, as before, that the error terms in the selection equation $v_{it} = \xi_i + \eta_{it}$ exhibit the usual one-way error components structure, with normally distributed components. That is

$$v_i | X_i \sim \text{NID}(\sigma_\xi^2 u' + \sigma_\eta^2 I) \; .$$

Note that while we do make explicit distributional assumptions about the disturbances in the main equation we assume

$$E(u_{it} | X_i, v_i) = \tau_1 v_{it} + \tau_2 \bar{v}_i \; . \tag{12.43}$$

Equation (12.43) implies that the conditional expectation $E(\varepsilon_{it} | X_i, z_{i0}, z_i)$ is a linear function of the conditional expectation of v_{it} and its individual specific mean noting that the τ's are parameters to be estimated. To derive the conditional expectation of the terms on the right hand side of (12.43) we use

$$E(u_{it} | X_i, z_{i0}, z_i) = \int [\xi_i + E(\eta_{it} | X_i, z_{i0}, z_i, \xi_i)] \, f(\xi_i | X_i, z_{i0}, z_i) \, d\xi_i \; , \tag{12.44}$$

where $f(\xi_i | X_i, z_{i0}, z_i)$ is the conditional density of ξ_i. The conditional expectation $E(\eta_{it} | X_i, z_{i0}, z_i, \xi_i)$ is the usual cross-sectional generalized residual (see Gourieroux et al., 1987, Vella, 1993) from (12.37) and (12.38), since, conditional on ξ_i, the errors from this equation are independent across observations. The conditional distribution of ξ_i can be derived using the result

$$f(\xi_i | X_i, z_{i0}, z_i) = \frac{f(z_i, z_{i0} | X_i, \xi_i) f(\xi_i)}{f(z_i, z_{i0} | X_i)} \; , \tag{12.45}$$

where we have used that ξ_i is independent of X_i and

$$f(z_i, z_{i0} | X_i) = \int f(z_i, z_{i0} | X_i, \xi_i) f(\xi_i) \mathrm{d}\xi_i \tag{12.46}$$

is the likelihood contribution of individual i in (12.37) and (12.38). Finally

$$f(z_i, z_{i0} | X_i, \xi_i) = \left[\prod_{t=1}^{T} f(z_{it} | X_i, z_{i,t-1}, \xi_i) \right] f(z_{i0} | X_i, \xi_i) \; , \tag{12.47}$$

where $f(z_{it} | X_i, z_{i,t-1}, \xi_i)$ has the form of the likelihood function in the cross-sectional case. If we assume that $f(z_{i0} | X_i, \xi_i)$ does not depend on ξ_i, or any of the other error components, then z_{i0} is exogenous and $f(z_{i0} | X_i, \xi_i) = f(z_{i0} | X_i)$. Thus we can condition on z_{i0} in (12.46) and (12.47) and obtain valid inferences neglecting its distribution. In general, however, we require an expression for the distribution of the initial value conditional on the exogenous variables and the ξ_i. As stated above

in the discussion of the random effects tobit model, the typical manner in which this is done is to follow Heckman (1981) in which the reduced form for z_{i0} is approximated using all presample information on the exogenous variables.[6]

Thus the two-step procedure takes the following form. The unknown parameters in (12.37) and (12.38) are estimated by maximum likelihood while exploiting the random effects structure. Equation (12.44) is then evaluated at these ML estimates by employing the expression for the likelihood function in an i.i.d. context, the corresponding generalized residual, and the numerical evaluation of two one dimensional integrals. This estimate, and its average over time for each individual provide two additional terms to be included in the primary equation. The additional parameters, corresponding to τ_1 and τ_2, can then be estimated jointly with β. Under the null hypothesis of exogenous sample selection, $\tau_1 = \tau_2 = 0$, and there is no need to adjust the standard errors. Thus the standard Wald test is a test of sample selection bias. However, in general the standard errors need to be adjusted for heteroskedasticity, serial correlation, and for estimation of the correction terms.

As we noted above, the correction terms have been written to allow greater flexibility with respect to the censoring process. We address this issue in the following section. However, as the model in (12.36), (12.37) and (12.38) is perhaps the most commonly encountered for panel data models with selectivity it is useful to see the form of the correction terms. The first step is to estimate the model by random effects probit to obtain estimates of the $\theta_2's$ and the variances σ_ξ^2 and σ_η^2. We then compute (12.44) and its individual specific average after inserting the following terms

$$E(\eta_{it}|X_i, z_{i0}, z_i, \xi_i) = d_{it}\sigma_\eta \frac{\Phi\left(\dfrac{x_{it}'\theta_{21} + x_{1it}'\theta_{22} + \xi_i}{\sigma_\eta}\right)}{\Phi\left(d_{it}\dfrac{x_{it}'\theta_{21} + x_{1it}'\theta_{22} + \xi_i}{\sigma_\eta}\right)}, \tag{12.48}$$

where Φ denotes the standard normal density function, and

$$f(\xi_i|X_i, z_{i0}, z_i) = \frac{\prod_{t=1}^T \Phi\left(d_{it}\dfrac{x_{it}'\theta_{21} + x_{1it}'\theta_{22} + \xi_i}{\sigma_\eta}\right)\dfrac{1}{\sigma_\xi}\Phi\left(\dfrac{\xi_i}{\sigma_\xi}\right)}{\int_{-\infty}^{\infty}\prod_{t=1}^T \Phi\left(d_{it}\dfrac{x_{it}'\theta_{21} + x_{1it}'\theta_{22} + \xi}{\sigma_\eta}\right)\dfrac{1}{\sigma_\xi}\Phi\left(\dfrac{\xi}{\sigma_\xi}\right)\mathrm{d}\xi}, \tag{12.49}$$

where $d_{it} = 2z_{it} - 1$.

The model can be estimated by maximum likelihood if we make some additional distributional assumptions regarding the primary equation errors. If all error components are assumed to be homoskedastic and jointly normal, excluding autocorrelation in the time-varying components, it follows that (12.43) holds with $\tau_1 = \sigma_{\varepsilon\eta}/\sigma_\varepsilon^2$ and $\tau_2 = T(\sigma_{\alpha\xi} - \sigma_{\varepsilon\eta}\sigma_\xi^2/\sigma_\varepsilon^2)/(\sigma_\eta^2 + T\sigma_\xi^2)$. This shows that τ_2 is nonzero even when the individual effects α_i and ξ_i are uncorrelated. In contrast, the two-step approach readily allows for heteroskedasticity and autocorrelation in the

[6] Note that apart from its dependence on ξ_i, the specification of $f(z_{i0}|X_i, \xi_i)$ can be tested separately from the rest of the model.

primary equation. Moreover, the assumption in (12.43) can easily be relaxed to, for example:

$$E(u_{it}|X_i, v_i) = \lambda_{1t} v_{i1} + \lambda_{2t} v_{i2} + \ldots + \lambda_{Tt} v_{iT} \; . \tag{12.50}$$

By altering (12.43) this approach can be extended to multiple sample selection rules. With two selection rules, $z_{1,it}$ and $z_{2,it}$, say, with reduced form errors $v_{1,it}$ and $v_{2,it}$, respectively, (12.43) is replaced by

$$E(u_{it}|X_i, v_{1,i}, v_{2,i}) = \tau_{11} v_{1,it} + \tau_{12} \bar{v}_{1,i} + \tau_{21} v_{2,it} + \tau_{22} \bar{v}_{2,i} \; . \tag{12.51}$$

Computation of the generalized residuals, however, now requires the evaluation of $E\{v_{j,it}|X_i, z_{1,i}, z_{2,i}\}$ for $j = 1,2$. Unless $z_{1,i}$ and $z_{2,i}$ are independent, conditional upon X_i, the required expressions are different from those obtained from (12.44) and (12.45) and generally involve multi-dimensional numerical integration.

While the two-step procedure is highly parameterized, many of the above assumptions can be tested empirically. While relaxing normality in the reduced form is typically computationally difficult, it is possible to test for departures from normality. It is also possible to test for serial correlation and heteroskedasticity using conditional moment tests. Also, we assume that the variables in X_i are strictly exogenous. This assumption excludes lagged dependent variables in the primary equation as well as feedback from lagged values of y to current x's. If components of X_i are not strictly exogenous they should be included in z and excluded from the reduced form.

12.6.3 Alternative Selection Rules

The discussion above frequently assumes that the selection rule is based on a binary outcome and for this reason the selection process was based on the use of the random effects probit likelihood function. However, just as in the cross-sectional case where the selection rule has been extended to alternatives rather than just the binary case (see Vella, 1993) it is useful to do so in the panel context. Two obvious, and practical, extensions are the two following models. The first is the extension to panel data of the Tobit type 3 model given by

$$\begin{aligned}
y_{it}^* &= x_{it}' \beta_1 + \beta_2 z_{it} + u_{it}, \\
z_{it}^* &= x_{it}' \theta_{21} + x_{1it}' \theta_{22} + v_{it}, \\
z_{it} &= z_{it}^* \cdot I(z_{it}^* > 0), \\
y_{it} &= y_{it}^* \cdot I(z_{it}^* > 0).
\end{aligned}$$

In this case one sees that the primary equation may or may not have the censoring variable as an endogenous explanatory variable and the censoring equation is censored at zero but observed for positive values. In our wage example discussed above, the extension implies that we observe not only whether the individual works but also the number of hours. We also allow the number of hours to affect the wage rate. For

this model we would first estimate the censoring equation by random effects tobit. We would then use these estimates, along with the appropriate likelihood contribution and tobit generalized residual, to compute (12.44) which are to be included in the main equation. Note that due to the structure of the model the inclusion of the correction terms accounts for the endogeneity of z_{it} in the main equation.

A second model of interest is where the z_{it} is observed as an ordinal variable, taking values j for $j = 1,\ldots,J$, and where the values of y_{it} are only observed for certain values of j. In this case, where the dummies denoting the value of z_{it} do not appear in the model, we would conduct estimation in the following way. Estimate the censoring equation by random effects ordered probit and then compute the corrections based on (12.44) accordingly. Then estimate the main equation over the subsample for z_{it} corresponding to a specific value and including the correction terms. When one wishes to include the dummies denoting the value of z_{it} as additional explanatory variable it is necessary to pool the sample for the different values of z_{it} and include the appropriate corrections.

12.6.4 Two-Step Estimators with Fixed Effects

A feature of the two-step estimator discussed above is their reliance on the assumption that the individual effect is random variable and independent of the explanatory variables. While the approach proposed by Vella and Verbeek (1999) is somewhat able to relax the latter assumption it is generally difficult to overcome. For this reason, as we noted above in the discussion of the censored regression model, it is generally more appealing to treat the individual fixed component of the error term as a fixed effect which may be correlated with the explanatory variables. We noted above that the results of Hahn and Newey (2004) would allow one to estimate a fixed effects tobit model and then perform the appropriate bias correction. Accordingly, it would be useful to adopt the same approach in the sample selection model and this has been studied by Fernandez-Val and Vella (2005). The basic model they study has the form

$$y_{it}^{*} = x_{it}'\beta + \alpha_i + \varepsilon_{it}, \tag{12.52}$$

$$z_{it}^{*} = x_{it}'\theta_{21} + x_{1it}'\theta_{22} + \xi_i + \eta_{it}, \tag{12.53}$$

$$z_{it} = I(z_{it}^{*} > 0), \tag{12.54}$$

$$y_{it} = y_{it}^{*} \cdot z_{it}, \tag{12.55}$$

where the α_i and ξ_i are individual specific fixed effects, potentially correlated with each other and the explanatory variables, and the ε_{it} and ξ_{it} are random disturbances which are jointly normally distributed and independent of the explanatory variables. While Fernandez-Val and Vella (2005) consider various forms of the censoring function, such as described in the previous section, we focus here on the standard case where the selection rule is a binary censoring rule.

The estimators proposed by Fernandez-Val and Vella (2005) are based on the following approach. One first estimates the reduced form censoring rule by the

appropriate fixed effects procedure. In the case of the standard selection rule this would be fixed effects probit but in the case of tobit censoring rules or ordered selection rules one would then use fixed effects tobit or fixed effects ordered probit respectively. Once these estimates are obtained one uses the bias correction approaches outlined in Hahn and Newey (2004) to adjust the estimates. With these bias corrected estimates one then computes the appropriate correction terms which generally correspond to the cross-sectional generalized residuals. One then estimates the main equation, (12.52), by a linear fixed effects procedure and bias correct the estimates. Fernandez-Val and Vella (2005) study the performance of this procedure to a range of models for alternative forms of censoring. These include the static and dynamic binary selection rule, and the static and dynamic tobit selection rule. They find that the Monte Carlo evidence suggests these procedures are very effective in eliminating selectivity bias. In instances where the adjustments were made to account for the endogeneity of the explanatory variables the procedures were also effective.

12.6.5 Semi-parametric Sample Selection Models

Kyriazidou (1997) also studied the model in (12.52), (12.53), (12.54) and (12.55). Her approach is semi-parametric in the sense that no assumptions are placed on the individual specific effects α_i and ξ_i and the distributional assumptions on the transitory errors ε_{it} and η_{it} are weak.

It is clear that $(\theta_{21}, \theta_{22})$ can be estimated by one of the methods for estimation of discrete choice models with individual specific effects, such as Rasch's (1960, 1961) conditional maximum likelihood estimator, Manski's (1987) maximum score estimator or the smoothed versions of the conditional maximum score estimator considered in Charlier et al. (1995) or Kyriazidou (1995). Kyriazidou's insight into estimation of β combines insights from the literature on the estimation of semi-parametric sample selection models (see Powell, 1987) with the idea of eliminating the individual specific effects by differencing the data. Specifically, to difference out the individual specific effects α_i, one must restrict attention to time periods s and t for which y is observed. With this "sample selection", the mean of the error term in period t is

$$\lambda_{it} = E(\varepsilon_{it} | \eta_{it} > -x'_{it}\theta_{21} - x'_{1it}\theta_{22} - \xi_i, \eta_{is} > -x'_{is}\theta_{21} - x'_{1is}\theta_{22} - \xi_i, \zeta_i)$$

where $\zeta_i = (x_{is}, x_{1is}, x_{it}, x_{1it}, \alpha_i, \xi_i)$. The key observation in Kyriazidou (1997) is that if $(\varepsilon_{it}, \eta_{it})$ and $(\varepsilon_{is}, \eta_{is})$ are independent and identically distributed (conditional on $(x_{is}, x_{1is}, x_{it}, x_{1it}, \alpha_i, \xi_i)$), then for an individual i, who has $x'_{it}\theta_{21} + x'_{1it}\theta_{22} = x'_{is}\theta_{21} + x'_{1is}\theta_{22}$,

$$\begin{aligned}\lambda_{it} &= E(\varepsilon_{it} | \eta_{it} > -x'_{it}\theta_{21} - x'_{1it}\theta_{22} - \xi_i, \zeta_i) \\ &= E(\varepsilon_{is} | \eta_{is} > -x'_{is}\theta_{21} - x'_{1is}\theta_{22} - \xi_i, \zeta_i) \\ &= \lambda_{is}. \end{aligned} \tag{12.56}$$

This implies that for individuals with $x_{it}'\theta_{21} + x_{1it}'\theta_{22} = x_{is}'\theta_{21} + x_{1is}'\theta_{22}$, the same differencing that will eliminate the fixed effect will also eliminate the effect of sample selection. This suggests a two-step estimation procedure similar to Heckman's (1976, 1979) two-step estimator of sample selection models: first estimate $(\theta_{21}, \theta_{22})$ by one of the methods mentioned earlier, and then estimate β by applying OLS to the first differences, but giving more weight to observations for which $(x_{it} - x_{is})'\widehat{\theta}_{21} + (x_{1it} - x_{1is})\widehat{\theta}_{22}$ is close to zero:

$$\hat{\beta}_2 = \left[\sum_{i=1}^{n}\sum_{s<t}(x_{it}-x_{is})'(x_{it}-x_{is})K\left(\frac{(x_{it}-x_{is})'\widehat{\theta}_{21}+(x_{1it}-x_{1is})\widehat{\theta}_{22}}{h_n}\right)y_{it}y_{is}\right]^{-1}$$
$$\times\left[\sum_{i=1}^{n}\sum_{s<t}(x_{it}-x_{is})'(x_{it}-x_{is})K\left(\frac{(x_{it}-x_{is})'\widehat{\theta}_{21}+(x_{1it}-x_{1is})\widehat{\theta}_{22}}{h_n}\right)y_{it}y_{is}\right]$$

where K is a kernel and h_n is a bandwidth which shrinks to zero as the sample size increases. Kyriazidou (1997) showed that the resulting estimator is $\sqrt{n}$-consistent and asymptotically normal. Kyriazidou (2001) shows how the same approach can be used to estimate models when lagged dependent variables are included as explanatory variables in (12.52) or (12.53).

As pointed out in Honoré and Kyriazidou (2000), the estimators proposed in Honoré (1992) and Kyriazidou (1997) can be modified fairly trivially to cover static panel data versions of the other tobit-type models discussed in Amemiya (1985).

12.6.6 Semi-parametric Estimation of a Type-3 Tobit Model

One paper which explores the semi-parametric estimation of panel data models with a tobit type censoring rule is Lee and Vella (2006). To present this idea first consider the cross-sectional estimator they propose.[7] They consider the following model:

$$y_i = x_i'\beta + u_i, \tag{12.57}$$
$$z_i^* = x_{it}'\theta_{21} + x_{1it}'\theta_{22} + v_i \tag{12.58}$$
$$z_i = \max(0, z_i^*), \quad s_i = I(z_i > 0), \tag{12.59}$$
$$(x_i',\ z_i,\ s_i y_i)' \text{ is observed, i.i.d. across } i, \tag{12.60}$$

and impose the following mean independence assumption $E(u_i|v_i, x_i, s_i)$=$E(u_i|v_i, s_i)$. The approach to obtain consistent estimates of β is to purge the (12.57) equation of the component related to the selection equation (12.58) error. To do this they suggest a Robinson (1988) type procedure in which they regress $y_i - E(y_i|v_i, s_i = 1)$ on $x_i - E(x_i|v_i, s_i = 1)$ noting the inclusion of v_i in the conditioning set eliminates

[7] Semi-parametric estimation of the cross-sectional form of this model is also considered in Honoré et al. (1997), Chen (1997) and Li and Wooldridge (2002).

the source of the selection problem.[8] The model is semi-parametric in that one does not make distributional assumptions about the disturbances. Rather, one estimates the selection model (12.58) and (12.59) parameters by some appropriate semi-parametric estimator and the estimates $\widehat{v}_i$ as $z_i - x_{1i}\widehat{\theta}_{21} - x'_{2i}\widehat{\theta}_{22}$ (if $s_i = 1$), where the $\widehat{\theta}_{21}$ and $\widehat{\theta}_{22}$ denote the first step semi-parametric estimates. The expectations $E(y_i|v_i, s_i = 1)$ and $E(x_i|v_i, s_i = 1)$ can be estimated non-parametrically. Lee and Vella (2006) argue that this approach can be extended to additional forms of endogeneity and selectivity by simply including the appropriate reduced form residual(s) in the conditioning set. This type of estimator is useful in the two wave panel context and Lee and Vella consider two models which adopt alternative strategies for dealing with dynamics in the model. The first is where the lagged dependent variable appears in the conditional mean and the model has the following form:

$$\begin{aligned}
&y_{it} = y_{i,t-1}\beta_y + x'_{it}\beta + u_{it},\\
&z^*_{it} = x'_{it}\theta_{21} + x'_{1it}\theta_{22} + v_{it}\\
&z_{it} = \max(0, z^*_{it}), \quad s_{it} = I(z_{it} > 0), \ \ t = 1, 2,\\
&(x'_{i1}, x'_{i2},\ z_{i1}, z_{i2},\ s_{i1}y_{i1}, s_{i2}y_{i2})' \text{ is observed, i.i.d. across } i. \qquad (12.61)
\end{aligned}$$

The outcome equation can only be estimated over the subpopulation $s_{i1} = s_{i2} = 1$, which poses a double selection problem. Thus one estimates over this subsample after subtracting off the component of the outcome equation related to the two selection residuals. The mean independence condition assumption required is $E(u_{i2}|v_{i1}, v_{i2}, x_{i2}, y_{i1}, s_i) = E(u_{i2}|v_{i1}, v_{i2}, s_i)$ and one estimates

$$y_{i2} - E(y_{i2}|v_{i1}, v_{i2}) = [y_{i1} - E(y_{i1}|v_{i1}, v_{i2})]\beta_y + [x_{i2} - E(x_{i2}|v_{i1}, v_{i2})]'\beta + \varepsilon$$

over the subsample corresponding to $s_{i1} = s_{i2} = 1$. Lee and Vella also consider the treatment of dynamics through the inclusion of a time invariant individual fixed effect α_i. The main equation is static and is of the form:

$$y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it}\ .$$

The double selection problem arises if the first-differenced outcome equation is estimated to eliminate a time-constant error which is potentially related to x_{it}'s:

$$\Delta y_i = \Delta x'_i\beta + \Delta\varepsilon_i, \quad \Delta y_i \equiv y_{i2} - y_{i1},\ \Delta x_i \equiv x_{i2} - x_{i1},\ \Delta\varepsilon_i \equiv \varepsilon_{i2} - \varepsilon_{i1}\ .$$

The mean independence assumption required is $E(\Delta\varepsilon_i|v_{i1}, v_{i2}, \Delta x_i, s_i) = E(\Delta\varepsilon_i|v_{i1}, v_{i2}, s_i)$ and one estimates

$$\Delta y_i - E(\Delta y_i|v_{i1}, v_{i2}) = [\Delta x_i - E(\Delta x_i|v_{i1}, v_{i2})]'\beta + \varepsilon$$

over the subsample corresponding to $s_{i1} = s_{i2} = 1$.

[8] The same estimator for the cross-sectional case was independently suggested in Li and Wooldridge (2002).

12.7 Some Empirical Applications

We conclude this chapter by discussing several empirical papers in which special cases of the general model in (12.1), (12.2), (12.3) and (12.4) are implemented. In each of these applications, economic agents select themselves into a certain state (e.g. "working", "union member", or "participant in a social program") and this self-selection is likely to be endogenous. In most cases fully parametric estimators are employed.

12.7.1 Attrition in Experimental Data

Hausman and Wise (1979) was one of the first studies to discuss the problem of attrition bias in experimental or panel data. Their analysis was aimed at measuring the effects of the Gary income maintenance experiment. In this experiment people were exposed to a particular income/tax policy, and the effects of this policy on monthly earnings were studied. Their sample consisted of 585 black males observed before the experiment took place ($t = 1$). In the second period, a treatment (i.e., an income guarantee/tax rate combination) was given to 57% of them, the other part was kept in the sample as a control group. So to analyze the effects of the experiment, Hausman and Wise were able to compare the behavior of a treatment group with that of a contemporaneously observed control group, as well as with its own pre-experimental behavior. The problem with estimating the effects from the experiment on earnings was that the second period suffered from high rates of attrition. From the experimental group 31% dropped out of the sample, while almost 41% of the individuals in the control group were not observed at $t = 2$. Moreover, it was thought likely that those individuals stay in the sample that benefit most from the experiment, i.e., those individuals that experience an increase in earnings due to the experiment. Obviously, such a self-selection is related to the unobservables in the equation of interest, which makes a tobit-type model appropriate.

The model considered by Hausman and Wise (1979) is fairly simple, because it is a fully-parametric two period model, where attrition (self-selection) only takes place in the second period. For each individual a treatment dummy d_{it} is defined, which is equal to zero at $t = 1$ for all individuals and equal to one in period 2 for those individuals that receive treatment. The model is then given by

$$y_{it}^* = d_{it}\delta + x_{it}'\beta + \alpha_i + \varepsilon_{it},\ t = 1,2\,, \tag{12.62}$$

where δ measures the effect of the treatment ("the treatment effect") and where x_{it} contains individual-specific exogenous variables, including an intercept or a time trend. Because (by assumption) selection takes place in the second period only, the selection equation can be described by a univariate probit model. In particular, it is assumed that y_{it} is observed if $t = 1$ and if $z_{i2} = I(z_{i2}^* > 0) = 1$, where

$$z_{i2}^* = x_{i2}'\theta_{21} + x_{1i2}'\theta_{22} + y_{i2}\gamma + v_{i2}^* \,. \tag{12.63}$$

All error terms are assumed to be normally distributed, with mutual independence of α_i, ε_{it} and v_{i2}^*. Note that, unlike before, (12.62) is not written as a reduced form and includes the dependent variable from the primary equation. As long as $\gamma = 0$. attrition depends upon the endogenous variable y_{i2} and OLS applied to (12.62) is inconsistent. Because y_{i2} is not observed for those individuals with $z_{i2} = 0$, we substitute (12.62) into (12.62) to obtain the reduced form

$$z_{i2}^* = x_{i2}'\theta_{21} + x_{1i2}'\theta_{22} + (d_{i2}\delta + x_{i2}'\beta)\gamma + (\alpha_i + \varepsilon_{i2})\gamma + v_{i2}^* \,, \tag{12.64}$$

or, after some appropriate redefinitions,

$$z_{i2}^* = w_{i2}'\theta_2 + x_{1i2}'\theta_{22} + v_{i2} \,. \tag{12.65}$$

The probit error term $v_{i2} = (\alpha_i + \varepsilon_{i2})\gamma + v_{i2}^*$ will be correlated with both α_i and ε_{i2} as long as $\gamma \neq 0$. Consequently, if one selects on participation in period 2 ($z_{i2} = 1$), this may not only affect inferences for period 2, but also for period 1 (unless $\sigma_\alpha^2 = 0$).

The likelihood contributions of the model consisting of (12.62) and (12.65) are given in Hausman and Wise and are special cases of those considered in Sect. 12.6.1. If specification (12.62) contains a time effect and a treatment dummy only, OLS produces an estimate of the treatment effect of -0.06. Correcting for attrition bias and applying maximum likelihood increases this effect to -0.11. If (12.62) contains a number of additional explanatory variables, both approaches yield roughly the same answer: -0.08. Consequently, Hausman and Wise conclude that within the context of a structural model, some attrition bias seems to be present, but not enough to substantially alter the estimate of the experimental effect.

In the Hausman and Wise model, it is assumed that selection into the experiment is random. In many other cases, however, individuals are allowed to select themselves into the experiment. Even in the absence of attrition this may lead to a selection bias problem. See Heckman (2001) or Wooldridge (2002, Chap. 18) for more discussion.

12.7.2 Real Wages Over the Business Cycle

Keynes (1936) believed that the movement of real wages over the business cycle was countercyclical. A large number of empirical studies on this issue, based on macro as well as micro data, have lead to a diversity of results. In an attempt to reconcile these results, Keane et al. (1988) consider the question to what extent aggregation bias (or selection bias) is able to explain the differences. Aggregation bias arises if people going in and out of the labor force are not random. In that case, the average wage changes over time due to a changing composition of the work force, even though real wage levels are unaffected. If, for example, low-wage industries are more cyclically sensitive, a countercyclical bias in the conclusion is expected.

Keane et al. (1988) use panel data from the National Longitudinal Survey of Young Men (NLS) over the period 1966–1981. The use of micro data has the advantage that a large part of the individual heterogeneity is observed. The model is given by

$$y_{it}^{*} = x_{it}'\beta_1 + \beta_2 ur_t + \alpha_i + \varepsilon_{it} \ , \tag{12.66}$$

where y_{it}^{*} is the log of the (potentially unobserved) real hourly wage rate, and ur_t denotes the national unemployment rate. The vector x_{it} contains individual-specific variables (education, experience, race, etc.), as well as a time trend. The parameter β_2 is the main parameter of interest: a positive β_2 corresponds to a countercyclical behavior in the wage, while a negative value indicates procyclical behavior. To correct for the possibility of aggregation bias (selection bias), there is an additional equation explaining employment, given by

$$z_{it}^{*} = x_{it}'\theta_{21} + x_{1,it}'\theta_{22} + \xi_i + \eta_{it} \ . \tag{12.67}$$

An individual is employed (and a wage rate is observed) if $z_{it} = I(z_{it}^{*} > 0) = 1$. Thus we have

$$y_{it} = y_{it}^{*} \cdot I(z_{it}^{*} > 0) \ . \tag{12.68}$$

Aggregation bias is procyclical if the covariance between the error terms in (12.66) and (12.67) is negative. In that case, individuals with relatively high wages are more likely to leave the labor market in case of increasing employment.

Keane et al. (1988) estimate two different specification of the model: one excluding individual-specific variables in (12.66) and (12.67) and one including a small number of such variables. In addition, four different estimation strategies are used: OLS without any corrections, maximum likelihood without individual effects in (12.66) and (12.67), with random effects (along the lines discussed in Sect. 12.6.1) and with fixed effects. Where needed, normality of the error components is assumed. The OLS estimate for β_2 of -0.0071 shows evidence of a procyclical behavior in the wage. The addition of the extra regressors results in an estimate of -0.0096, implying that the failure to control for observed heterogeneity leads to a countercyclical bias. The estimates from the fixed effects model show insignificant unemployment rate coefficients, implying an acyclic wage. The correlation coefficient between η_{it} and v_{it} is estimated to be -0.222. This result implies that the OLS unemployment coefficient is procyclically biased. Finally, if a random effects specification is estimated, the unemployment rate coefficients are negative and significant in both specifications. For the specification including observed heterogeneity the unemployment rate coefficient of -0.0066 is still considerably below the OLS effect of -0.0096. This indicates that a procyclical bias is still present, but weaker than was indicated by the fixed effects model. The random effects results indicate a negative correlation of the transitory errors (the correlation coefficient between η_{it} and v_{it} is -0.252), but a positive correlation of the individual effects α_i and ξ_i (with an estimated correlation coefficient of 0.436). The resulting composite correlation is virtually zero.

The general conclusion from the results is that the failure to account for selection effects, biases the behavior of the real wage in a procyclical direction. Apparently, high-wage workers are more likely to become unemployed in a downturn.

12.7.3 Unions and Wages

Empirical studies of the union impact on wages typically attempt to estimate how observationally equivalent workers' wages differ in union and non-union employment. This is known as the "union effect". However, as the unobserved factors that influence the sorting into union and non-union employment may also affect wages it is necessary to incorporate how the unobserved heterogeneity responsible for the union/non-union decision is rewarded in the two sectors. Panel data studies of the union effect generally control for this endogeneity through fixed effects or alternative instrumental variables estimators. These procedures are inflexible in their treatment of worker heterogeneity as they generally assume the endogeneity is individual specific and fixed. A preferable approach, adopted by Vella and Verbeek (1998), is based on decomposing the endogeneity underlying union status into an individual specific component and an individual/time specific effect.

Vella and Verbeek (1998) consider the following equations explaining (log) union wages $y_{1,it}$ and (log) non-union wages $y_{0,it}$,

$$y_{j,it} = x'_{j,it}\beta_j + \alpha_{j,i} + \varepsilon_{j,it},\ j = 0,1\ , \tag{12.69}$$

where $x_{j,it}$ is a vector of characteristics, including time dummies. For a given individual, we observe his wage in the union or the non-union sector. Selection into the union sector is described by a random effects probit model of the form

$$\begin{aligned} z^*_{it} &= x'_{it}\theta_{21} + x'_{1,it}\theta_{22} + z_{i,t-1}\delta + \xi_i + \eta_{it} \\ z_{it} &= I(z^*_{it} > 0). \end{aligned} \tag{12.70}$$

This is a dynamic model, implying that the probability of working in the union sector is affected by the worker's status in the previous year.

The random components are assumed to be i.i.d. drawings from a multivariate normal distribution, where the effects from the different equations are potentially correlated. The endogeneity of union status (z_{it}) is driven by the correlations between the components in (12.69) and (12.70). The wage equation is estimated by OLS, fixed effects and by the two-step method described in Sect. 12.6.1, imposing that β_j is identical across sectors (except for the intercept term). The differences in the intercept capture the union effect. The data, taken from the National Longitudinal Survey (Youth Sample), comprise a sample of full-time working males in the USA, who have completed their schooling by 1980 and who are then followed over the period 1980–1987. This provides a balanced panel of 545 individuals.

The estimates for the union effect vary widely across the different methods. The OLS estimate is 0.146, corresponding to a union effect of about 15%. After correcting for the endogeneity of union status, the estimated union effect increases to 0.214 or about 21%. However, the random effects contribute significantly, making the union premium highly variable across individuals. Interestingly, the empirical results indicate that the random effects are valued differently by sector. That is, it is inappropriate to assume that the random components in (12.69) are identical for both sectors ($j = 0$ and $j = 1$). This is consistent with the idea that workers have sector-specific skills.

References

Amemiya, T. (1984): Tobit Models: A Survey, *Journal of Econometrics*, 24, 3–61.

Amemiya, T. (1985): *Advanced Econometrics*, Harvard University Press, Cambridge, MA.

Arellano, M., O. Bover and J.M. Labeaga (1999): Autoregressive Models with Sample Selectivity for Panel Data, in C. Hsiao, K. Lahiri, L. F. Lee, and H. Pesaran, eds., *Analysis of Panels and Limited Dependent Variable Models*, Cambridge University Press, Cambridge, MA, 23–48.

Charlier, E., B. Melenberg and A. van Soest (1995): A Smoothed Maximum Score Estimator for the Binary Choice Panel Data Model and an Application to Labour Force Participation, *Statistica Neerlandica*, 49, 324–342.

Charlier, E., B. Melenberg and A. van Soest (2000): Estimation of a Censored Regression Panel Data Model Using Conditional Moments Restrictions Efficiently, *Journal of Econometrics*, 95, 25–56.

Chen, S. (1997): Semi-Parametric Estimation of the Type 3 Tobit Model, *Journal of Econometrics,* 80, 1–34.

Fernandez-Val, I. and F. Vella (2005): Fixed Effects Estimation of Panel Models with Selection Bias, unpublished manuscript, MIT.

Gourieroux, C., A. Monfort, E. Renault and A. Trognon (1987): Generalized Residuals, *Journal of Econometrics*, 34, 5–32.

Greene, W.H. (2004): Fixed Effects and Bias Due to the Incidental Parameters Problem in the Tobit Model, *Econometric Reviews*, 23, 125–147.

Gronau, R. (1974): Wage Comparisons – A Selectivity Bias, *Journal of Political Economy*, 82, 1119–1143.

Hahn, J. and W.K. Newey (2004): Jacknife and Analytical Bias Reduction for Non-Linear Panel Models, *Econometrica*, 72, 1295–1319.

Hausman, J.A. and D.A. Wise (1979): Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment, *Econometrica*, 47, 455–473.

Heckman, J.J. (1974): Shadow Prices, Market Wages and Labor Supply, *Econometrica*, 42, 679–694.

Heckman, J.J. (1976): The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple estimator for Such Models, *Annals of Economic and Social Measurement*, 15, 475–492.

Heckman, J.J. (1979): Sample Selection Bias as a Specification Error, *Econometrica*, 47, 153–161.

Heckman, J.J. (1981): The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete-Time Discrete-Data Stochastic Process, in C.F. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, MA, 179–195.

Heckman, J.J. (2001): Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture, *Journal of Political Economy*, 109, 673–748.

Honoré, B.E. (1992): Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects, *Econometrica*, 60, 533–565.
Honoré, B.E. (1993): Orthogonality Conditions for Tobit Models with Fixed Effects and Lagged Dependent Variables, *Journal of Econometrics*, 59, 35–61.
Honoré, B.E. and L. Hu (2004): Estimation of Cross-Sectional and Panel Data Censored Regression Models with Endogeneity, *Journal of Econometrics*,122, 293–316.
Honoré, B.E. and E. Kyriazidou (2000): Estimation of Tobit–Type Models with Individual Specific Effects, *Econometric Reviews*, 19, 341–366.
Honoré, B., E. Kyriazidou and C. Udry (1997): Estimation of Type 3 Tobit Models Using Symmetric Trimming and Pairwise Comparisons, *Journal of Econometrics,* 76, 107–128.
Hu, L. (2002): Estimation of a Censored Dynamic Panel Data Model, *Econometrica*, 70, 2499–2517.
Keane, M., R. Moffitt and D. Runkle (1988): Real Wages Over the Business Cycle: Estimating the Impact of Heterogeneity with Micro Data, *Journal of Political Economy*, 96, 1232–1266.
Keynes, J.M. (1936): *The General Theory of Employment, Interest and Money*, MacMillan, London.
Kyriazidou, E. (1995): *Essays in Estimation and Testing of Econometric Models*, Northwestern University, Ph.D. dissertation.
Kyriazidou, E. (1997): Estimation of a Panel Data Sample Selection Model, *Econometrica*, 65, 1335–1364.
Kyriazidou, E. (2001): Estimation of Dynamic Panel Data Sample Selection Models, *Review of Economic Studies*, 68, 543–572.
Lancaster, T. (2000): The Incidental Parameter Problem Since 1948, *Journal of Econometrics*, 95, 391–413.
Lee, M.J. and F. Vella (2006): A Semi-Parametric Estimator for Censored Selection Models with Endogeneity, *Journal of Econometrics*, 130, 235–252.
Li, Q. and J.M. Wooldridge (2002): Semi-Parametric Estimation of Partially Linear Models for Dependent Data with Generated Regressors, *Econometric Theory,* 18, 625–645.
Manski, C. (1987): Semi-Parametric Analysis of Random Effects Linear Models from Binary Panel Data, *Econometrica*, 55, 357–362.
Newey, W.K. (1984): A Method of Moments Interpretation of Sequential Estimators, *Economics Letters*, 14, 201–206.
Neyman, J. and E. Scott (1948): Consistent Estimates Based on Partially Consistent Observations, *Econometrica*, 16, 1–32.
Nijman, T.E. and M. Verbeek (1992): Nonresponse in Panel Data: The Impact on Estimates of a Life Cycle Consumption Function, *Journal of Applied Econometrics*, 7, 243–257.
Powell, J.L. (1987): *Semi-Parametric Estimation of Bivariate Latent Models*, Working Paper No. 8704, Social Systems Research Institute, University of Wisconsin–Madison.
Quenouille, M. (1956): Notes on Bias in Estimation, *Biometrika,* 43, 353–360.
Rasch, G. (1960): *Probabilistic Models for Some Intelligence and Attainment Tests*, Denmarks Pædagogiske Institut, Copenhagen.
Rasch, G. (1961): On the General Laws and the Meaning of Measurement in Psychology, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, University of California Press, Berkeley and Los Angeles.
Ridder, G. (1990): Attrition in Multi-Wave Panel Data, in J. Hartog, G. Ridder and J. Theeuwes, eds., *Panel Data and Labor Market Studies*, Elsevier, North Holland.
Rivers, D. and Q. Vuong (1988): Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models, *Journal of Econometrics*, 39, 347–366.
Robinson, P. (1988), Root-*N* Consistent Semi-Parametric Regression, *Econometrica,* 56, 931–954.
Smith, R.J. and R. Blundell (1986): An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply, *Econometrica*, 54, 679–685.
Tukey, J. (1958): Bias and Confidence in Not-Quite Large Samples, *Annals of Mathematical Statistics,* 29, 614.

Vella, F. (1993): A Simple Estimator for Simultaneous Models with Censored Endogenous Regressors, *International Economic Review*, 34, 441–457.

Vella, F. (1998): Estimating Models with Sample Selection Bias: A Survey, *Journal of Human Resources,* 33, 127–169.

Vella, F. and M. Verbeek (1998): Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men, *Journal of Applied Econometrics*, 13, 163–189.

Vella, F. and M. Verbeek (1999): Two-Step Estimation of Panel Data Models with Censored Endogenous Regressors and Selection Bias, *Journal of Econometrics*, 90, 239–263.

Verbeek, M. and T.E. Nijman (1996): Incomplete Panels and Selection Bias, in L. Matyas and P. Sevestre, eds., *The Econometrics of Panel Data. Handbook of the Theory with Applications*, 2nd edition, Kluwer Academic Publishers, Dordrecht, 449–490.

Waterman, R., B. Lindsay, H. Li and B. Li (2000): Projected Score Methods for Nuisance Parameters: Asymptotics and Neyman-Scott Problems, unpublished manuscript.

Wooldridge, J.M. (1995): Selection Corrections for Panel Data Models Under Conditional Mean Independence Assumptions, *Journal of Econometrics*, 68, 115–132.

Wooldridge, J.M. (2002): *Econometric Analysis of Cross-Section and Panel Data*, MIT Press, Cambridge, MA.

Wooldridge, J.M. (2005): Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity, *Journal of Applied Econometrics*, 20, 39–54.

Woutersen, T.M. (2002): Robustness Against Incidental Parameters, unpublished manuscript, Johns Hopkins University.

Chapter 13
Simulation Techniques for Panels: Efficient Importance Sampling

Roman Liesenfeld and Jean-François Richard

13.1 Introduction

The recent years have witnessed an explosive growth in the use of simulation techniques in econometrics made possible by impressive advances in computing power. See e.g., the special issue of the *Journal of Applied Econometrics* Brown et al. (1993) on "Econometric Inference using Simulation Techniques". See also Stern (1997) for a survey of simulation-based estimation with special emphasis on multivariate probit models. Among the methods surveyed by Stern, one of particular interest is the popular GHK simulator developed by Geweke (1991), Hajivassiliou (1990) and Keane (1994) (see also Geweke et al. (1994) and Geweke et al. (1997)). Börsch–Supan and Hajivassiliou (1993) compare the performance of an acceptance-rejection method proposed by Stern (1992) with that of the GHK technique. See Börsch–Supan et al., (1990) for an empirical application of the GHK to a multiperiod-multinomial probit model of living arrangements for the elderly. Greene (2004) compares GMM estimations – as proposed by Bertschek and Lechner (1998) – with simulated maximum likelihood for panel probit models allowing for unobserved heterogeneity along the time dimension, with an application to product innovation activity of German manufacturing firms (initially studied by Bertschek (1995)). It is important to note that the simulation techniques proposed in these papers are typically of low dimensions (either time or decision space) but cannot easily deal with the much higher dimensions required to handle random effects across individual units. This chapter is not aimed at providing a complete survey of the relevant literature. Instead we will draw upon our own experience with the

Roman Liesenfeld
Department of Economics, Christian-Albrechts-Universität Kiel, Olshausenstr. 40-60, 24118 Kiel, Germany, e-mail: liesenfeld@stat-econ.uni-kiel.de

Jean-François Richard
Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260, USA,
e-mail: fantin@pitt.edu

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

use of Monte-Carlo (hereafter MC) simulation techniques in classical and Bayesian econometrics and will attempt to provide readers with an "helicopter tour"- of some of their key features.

Moreover, we will largely focus our attention on a powerful high dimensional simulation technique known as Efficient Importance Sampling (hereafter EIS) which, as we shall discuss, is particularly well adapted to handling unobserved heterogeneity in large panel data sets. (Another important simulation technique known as Gibbs sampling is presented in Chap. 15 and will, therefore, not be discussed here.) In particular, the full potential of EIS will be highlighted in the context of an empirical application to a panel logit model with random effects along both dimensions (time and individual). As we shall discuss further below, the EIS method allows for efficient evaluation of likelihood functions for non-linear panel data models in the presence of unobserved heterogeneity along both dimensions, where the interaction between individual and time random effects prevent factorizations into problems of lower dimensionality. This is an important extension for conceptional as well as practical reasons. By modeling unobserved heterogeneity across individual units in the form of random effects, we can identify the impact of time invariant regressors. On the practical side this possibility enables researchers to make informed choices among fixed and random effects formulations of their model, instead of being forced to select fixed effects by default.

The most important usage of MC-EIS techniques in modern econometrics is that of a powerful *numerical* technique for the evaluation of high–dimensional (functional) integrals which are analytically intractable. Therefore, in sharp contrast with recent literature, we will insist upon separate treatment of the numerical and statistical properties of "simulation estimators". In a fundamental sense, MC simulation estimators ought to be treated for what they are, i.e., *numerical* approximations to a set of statistics of interest (whose *statistical* properties are to be separately discussed and might themselves have to be numerically approximated).

This chapter is organized as follows: Random number generation is discussed in Sect. 13.2; Importance Sampling and EIS is introduced in Sect. 13.3; Simulation based inference techniques are surveyed in Sect. 13.4 with reference to panel data models; their numerical properties are analyzed in Sect. 13.5; An empirical application of EIS is presented in Sect. 13.6; Sect. 13.7 concludes; A technical appendix details the implementation of EIS for large panel data sets up to a degree of details allowing for customization to one's specific application.

As for notation, we will use matched pairs of capital and lower case letters to respectively denote random variables and realizations thereof. A superscript $\sim$ denotes random draws. A superscript – paired with a subscript S denotes arithmetic sample means over S random draws.

13.2 Pseudorandom Number Generation

The cornerstone of any simulation based inference technique lies in the generation by computer of sequences of "pseudorandom" numbers. Following Devroye (1986), we will assume the availability of a uniform $[0,1]$ (linear congruential)

pseudorandom number generator. See Fishman and Moore (1982, 1986) for a statistical evaluation of congruential generators. Crude generators can be improved upon by "(re)shuffling", an option which is available in program libraries such as IMSL (see the 1991 IMSL User's Manual Stat/Library). See Press et al. (1986) for a portable reshuffling subprogram in FORTRAN or C. See also Fishman (1996) for a broad in-depth discussion of Monte Carlo simulation.

Sequences of pseudorandom numbers are in fact fully deterministic and uniquely characterized by their length and initial "seed". This fundamental property enables us to reproduce at will any sequence of pseudorandom numbers. It follows that regeneration of large sequences of pseudorandom numbers provides an efficient alternative to their storage on disk since (re)generation typically is faster than swapping. Also regeneration is the key to the technique of Common Random Numbers (CRN's) which plays an essential role in the design of simulation based inference techniques and is discussed in Sect. 13.3.3 below.

Uniform pseudorandom numbers can be transformed into pseudorandom draws from a broad range of distributions following techniques discussed, e.g. in Devroye (1986), the most important of which are surveyed here.

13.2.1 Univariate Distributions

Let X denote a univariate random variable with distribution function $F(x \mid \underline{\theta})$, where $\underline{\theta}$ is a vector of preassigned parameters. For the ease of presentation we will assume that X is a continuous random variable. Extensions to discrete random variables are straightforward. At a high level of generality pseudorandom draws of X are obtained by transformation of a sequence of independent uniform $[0,1]$ pseudorandom draws $\underline{U}' = (U_1, \ldots, U_K)$

$$X = \xi\left(\underline{U};\ \underline{\theta}\right), \tag{13.1}$$

where K is determined by an appropriate stopping rule and may itself be random. The most commonly used (transformation) techniques are inversion, rejection (or acceptance) and decomposition which are briefly described next. Additional details and extensions, as well as other techniques can be found in Devroye (1986).

13.2.1.1 Inversion

The random variable $U = F(X \mid \underline{\theta})$ is uniformly distributed on $[0,1]$. Let $F^{-1}(\cdot;\ \underline{\theta})$ denote the inverse of F for any given $\underline{\theta}$. It follows that the random variable

$$X = F^{-1}\left(U;\ \underline{\theta}\right) \tag{13.2}$$

when U is uniform on $[0,1]$ has F for its distribution function. The technique of inversion consists in generating sequences of uniform $[0,1]$ pseudorandom draws $\{U_i\}$ and transforming them into sequences $\{X_i\}$ by means of (13.2).

Example 13.1. Let X be a Weibull random variable with parameters $\alpha > 0$ and $\beta > 0$. Its distribution function is

$$F(x \mid \alpha, \beta) = 1 - \exp(-\beta x^{\alpha}) \ . \tag{13.3}$$

The transformation (13.2) is then given by

$$X = \left[-\frac{1}{\beta} \ln(1-U) \right]^{1/\alpha} \tag{13.4}$$

where $V = 1 - U$ is uniform on $[0,1]$.

The inversion technique is easy to implement as long as an analytical expression is available for F^{-1}. However, it is generally inefficient in terms of computing time relatively to more performant generation techniques combining fast rejection and decomposition as described below. This is particularly relevant if F has to be numerically inverted, in which case large scale simulation can become prohibitively time consuming. In such cases we recommend initial tabulation of the inverse for the relevant range of parameter values and subsequent numerical interpolation within the core simulation algorithm. This is quite relevant since, as we shall discuss further in Sect. 13.3.3 below, inversion though inefficient may be required in order to apply the critical concept of Common Random Numbers.

13.2.1.2 Rejection (Acceptance)

Assume X has a density function $f(x \mid \underline{\theta})$ which is not easily amenable to "direct" simulation (or for which direct simulation is relatively inefficient). For example, Bayesians and also users of Gibbs sampling techniques (as described in Chap. 15) frequently face situations where f is known only up to a proportionality factor. In such cases $f(x \mid \underline{\theta})$ is given in the form of a "density kernel" $\varphi(x \mid \underline{\theta})$ whose "integrating constant"

$$k(\underline{\theta}) = \int_{-\infty}^{\infty} \varphi(x \mid \underline{\theta}) \, \mathrm{d}x \tag{13.5}$$

requires numerical evaluation. The rejection technique which does not require knowledge of $k(\underline{\theta})$ is well adapted to such circumstances. It requires the construction of an "envelope" density function $\mu(x \mid \underline{\theta})$ for which a generator is available and which is such that

$$\sup_{x} \left[\frac{\varphi(x \mid \underline{\theta})}{\mu(x \mid \underline{\theta})} \right] = c(\underline{\theta}) < \infty \tag{13.6}$$

for all relevant $\underline{\theta}$'s. If, in particular, the support of μ is infinite, condition (13.6) requires that the tails of μ be not "thinner" than those of φ. It follows that normal density functions with their "thin" tails cannot be used as envelope densities for distributions with thicker tails such as student-t distributions, unless appropriate truncation of the support is deemed acceptable.

Under condition (13.6) the rejection technique runs as follows: to generate a random draw of X,

(a) generate a uniform U on $[0,1]$;
(b) generate Z with density function $\mu(z \mid \underline{\theta})$;
(c) If $c(\underline{\theta}) \cdot \mu(Z \mid \underline{\theta}) \cdot U > \varphi(Z \mid \underline{\theta})$, then "reject" Z and return to step (a); otherwise
(d) $X = Z$ (i.e., Z is "accepted"). It is straightforward to verify that

$$Pr(Z \leq a \mid Z \text{ is accepted}) = F(a \mid \underline{\theta}) \tag{13.7}$$

i.e., that X has the required distribution. The unconditional probability of acceptance is given by

$$Pr(Z \text{ is accepted}) = k(\underline{\theta}) \cdot [c(\underline{\theta})]^{-1} \leq 1 \,. \tag{13.8}$$

Devroye (1986) – see also Geweke (1994) – proposes to select optimized envelope density functions within a parametric class $M = \{m(x \mid \underline{\alpha}) ;\ \underline{\alpha} \in A\}$ by solving the following optimization problem

$$\underline{\alpha}^*(\underline{\theta}) = \arg \min_{\underline{\alpha} \in A} \left\{ \sup_x [\ln \varphi(x \mid \underline{\theta}) - \ln m(x \mid \underline{\alpha})] \right\} \tag{13.9}$$

and choosing $\mu(x \mid \underline{\theta}) = m(x \mid \underline{\alpha}^*(\underline{\theta}))$. The most performant generation algorithms for common distributions often use rejection in combination with the decomposition technique which is described next. Compound acceptance rates in excess of 0.95 are not uncommon. See Devroye, (1986) for examples and details.

We conclude this brief discussion of the rejection technique with two additional comments. First, the application of the rejection principle greatly simplifies if the random variable X can be obtained by means of a transformation $X = \delta(X_*; \underline{\theta})$ of a "standardized" random variable X_* with a density kernel φ_* which does not depend on $\underline{\theta}$. Under such circumstances we only have to construct a single optimized envelope function μ_* for φ_*, notwithstanding the fact that the construction of Common Random Numbers simplifies accordingly.

Second, the rejection algorithm as described above requires the evaluation of $\varphi(z \mid \underline{\theta})$ for all draws, which can be computationally demanding. A performant refinement of rejection is provided by the "squeeze" principle whose object is to squeeze $\varphi(z \mid \underline{\theta})$ between a pair of functions $a(z \mid \underline{\theta})$ and $b(z \mid \underline{\theta})$ which are (much) quicker to evaluate than φ itself (being, for example, piecewise linear). The squeeze version of step (c) of the rejection algorithm runs as follows:

(c.1) $c(\underline{\theta}) \cdot \mu(Z \mid \underline{\theta}) \cdot U < a(Z \mid \underline{\theta}) \Rightarrow$ "quick" acceptance of Z; else,
(c.2) $c(\underline{\theta}) \cdot \mu(Z \mid \underline{\theta}) \cdot U > b(Z \mid \underline{\theta}) \Rightarrow$ "quick" rejection of Z; else,
(c.3) run step (c) of the rejection algorithm.

Elaborate implementations of the squeeze principle may require evaluation of φ for less than 10% of the actual draws.

13.2.1.3 Decomposition

A powerful technique for the construction of highly efficient generators consists in the "decomposition" of a density kernel $\varphi(x \mid \underline{\theta})$ into a mixture of the form

$$\varphi(x \mid \underline{\theta}) = \sum_{i=1}^{k-1} p_i(\underline{\theta})\, q_i(x \mid \underline{\theta}) + p_k(\underline{\theta})\, q_k(x \mid \underline{\theta}) \ , \tag{13.10}$$

where the q_i's for $i : 1 \to k-1$ correspond to distributions that are obtained by simple transformations of uniforms on $[0,1]$, such as sums or extremums of pairs of uniforms. The q_i's often have non overlapping supports allowing for separate efficient treatments of the central and tail areas of φ. The "remainder" density kernel q_k typically requires more demanding generation procedures but efficient decompositions result in low values for the remainder probability $p_k(\underline{\theta})$. According to formula (13.10) a draw of X proceeds in two steps:

(a) A discrete random indicator $I \in \{1, 2, \ldots, k\}$ is drawn according to the probabilities $\{p_i(\underline{\theta})\}$;
(b) Conditionally on $I = i$, X is drawn from the distribution with kernel $q_i(x \mid \underline{\theta})$.

A continuous version of formula (13.10) is given by

$$\varphi(x \mid \underline{\theta}) = \int p(y \mid \underline{\theta})\, q(x \mid y,\, \underline{\theta})\, \mathrm{d}y \tag{13.11}$$

and can be used for such distributions as the student-t distribution (which is obtained by a continuous mixture of a Normal distribution for x conditionally on $y = \sigma^{-2}$ and a gamma distribution for σ^{-1}).

13.2.2 Multivariate Distributions

The generation of a multivariate random variable $\underline{X}$ is based upon recursive factorizations of its density kernel $\varphi(\underline{x} \mid \underline{\theta})$ into lower dimensional density kernels. Two types of factorizations are discussed next: sequential (or recursive) factorizations and Gibbs factorizations. Factorizations of different types often are combined together for the purpose of constructing a complete factorization of a multivariate distribution into univariate components to which the techniques described in Sect. 13.2.1 can be applied.

13.2.2.1 Sequential Factorizations

Sequential factorizations are typically used when the components of $\underline{X} = (\underline{X}_1, \ldots, \underline{X}_p)$ are naturally (pre)-ordered, such as in the context of time series problems. The density kernel of $\underline{X}$ is then factorized as

$$\varphi\left(\underline{x}_1, \ldots, \underline{x}_p \mid \underline{\theta}\right) = \prod_{i=1}^{p} \varphi_i(\underline{x}_i \mid \underline{x}_{(i-1)}, \underline{\theta}) \,, \tag{13.12}$$

where $\underline{x}_{(0)} = \phi$ and $\underline{x}'_{(i-1)} = (\underline{x}'_1, \ldots, \underline{x}'_{i-1})$ for $i : 2 \to p$. For the ease of notation preassigned (fixed) initial conditions are included in $\underline{\theta}$ and randomized initial conditions in $\underline{X}$ itself.

Let $\underline{\tilde{x}}$ denote an arbitrary random draw of $\underline{X}$. It is produced by sequential draws of the $\underline{X}_i$'s conditionally on $\underline{\theta}$ and $\underline{X}_{(i-1)} = \underline{\tilde{x}}_{(i-1)}$ according to the density kernels $\varphi_i(\underline{x}_i \mid \underline{\tilde{x}}_{(i-1)}, \underline{\theta})$ for $i : 1 \to p$. In the context of Gourieroux and Monfort (1994), to which we shall refer further below, such simulations are called "path simulations". Note that in the context of time series, models are generally formulated in sequential form and the joint density kernel $\varphi(\underline{x} \mid \underline{\theta})$ is rarely explicitely given. In other contexts the ordering of the components of $\underline{X}$ may be arbitrary and its actual choice based upon considerations of numerical convenience. For example, random draws of a multivariate Normal density are typically based upon a Cholesky decomposition of its covariance matrix for whatever ordering of its components is deemed to be convenient.

As discussed further below, conditional independence assumptions play a central role in the formulation of panel data models and can produce significant simplifications of formula (13.12). For example, in the context of the application discussed in Sect. 13.6 below, $\underline{x}$ is partitioned into $\underline{x}' = (\underline{\lambda}', \underline{\alpha}')$ where $\underline{\lambda}' = (\lambda_1, \ldots, \lambda_T)$ denotes time random effects and $\underline{\alpha}' = (\alpha_1, \ldots, \alpha_N)$ individual random effects in a non–linear panel model. Efficient numerical integration of $\underline{x}$ conditional on $\underline{\theta}$ *and* the actual sample $\underline{y}$ will be based upon a factorization of the form

$$\varphi\left(\underline{\lambda}, \underline{\alpha} \mid \underline{\theta}, \underline{y}\right) = \prod_{t=1}^{T} \varphi_t(\lambda_t \mid \underline{\lambda}_{(t-1)}, \underline{\alpha}, \underline{\theta}) \cdot \prod_{i=1}^{N} \varphi_i(\alpha_i \mid \underline{\theta}, \underline{\lambda}) \,. \tag{13.13}$$

Along similar lines, the concept of "exchangeable" distribution which is familiar to Bayesians assumes that the x_i's are identically independently distributed with density kernel q, conditionally on an unobserved common random component x_0 with density kernel p. Therefore, it is based upon a factorization of the form

$$\varphi\left(\underline{x}, \underline{\theta}\right) = \int p\left(x_0 \mid \underline{\theta}\right) \prod_{i=1}^{P} q\left(\underline{x}_i \mid x_0, \underline{\theta}\right) \mathrm{d}x_0 \,. \tag{13.14}$$

13.2.2.2 Gibbs Sampling

Gibbs sampling is extensively analyzed in Chap. 15 and is only briefly discussed here for the sake of completeness and comparison. Gibbs sampling is based upon the observation that a density kernel for the distribution of an arbitrary component $\underline{X}_i$ conditional on $\underline{\theta}$ and on *all* other components of $\underline{X}$ is trivially obtained by regrouping all terms depending upon $\underline{x}_i$, in the expression of $\varphi(\underline{x} \mid \underline{\theta})$. Depending upon the circumstances, such density kernels can either characterize

known distributions for which generators immediately are available or, in any event, are amenable to simulation by application of the principles described in Sect. 13.2.1.

Gibbs sampling is then based upon a Markov scheme as described in Chap. 15. It offers the advantage that it is easily implementable for a broad range of applications. It can, however, be inefficient as a large number of auxiliary draws are required in order to produce draws from the actual distribution of $\underline{X} \mid \underline{\theta}$.

13.3 Importance Sampling

While MC techniques have long been used to simulate the finite sample properties of a broad range of ("classical") statistics, one of the most important numerical development in recent years has been their increasing usage as a *numerical* method for evaluating (large-dimensional) analytically intractable integrals. It is hardly surprising that the initial impetus and many important developments on that front came from Bayesian statisticians and econometricans who critically depend upon the availability of efficient numerical procedures in order to evaluate posterior moments and other quantities of interest to them. See the pioneering contribution of Kloek and van Dijk (1978), or Geweke (1989, 1994) for more recent developments. The presentation which follows is based upon Richard and Zhang (2007). See also Liesenfeld and Richard (2003a) for a non technical presentation of Importance Sampling and Liesenfeld and Richard (2003b,c) for applications of EIS within the context of stochastic volatility models – including an explicit comparison with Gibbs sampling.

13.3.1 General Principle

Assume one has to evaluate a functional integral of the form

$$G(\underline{\delta}) = \int_{S(\underline{\delta})} g(\underline{x},\, \underline{\delta}) \cdot p(\underline{x} \mid \underline{\delta})\, \mathrm{d}\underline{x}\,, \tag{13.15}$$

where g is a function which is integrable *w.r.t.* a conditional density $p(\underline{x} \mid \underline{\delta})$ with support $S(\underline{\delta})$. The actual composition of $\underline{x}$ and $\underline{\delta}$ largely is problem dependent. Let $\underline{\theta}$, $\underline{y}$ and $\underline{\lambda}$ denote parameters, data and latent (unobservable) variables, respectively. Important applications requiring the evaluation of integrals of the form given is (13.15) are: (i) The Bayesian evaluation of posterior "odds" and/or moments for which $\underline{\delta} = \underline{y}$ and $\underline{x} = (\underline{\theta}, \underline{\lambda})$; (ii) The classical evaluation of "marginalized" likelihood functions for which $\underline{\delta} = (\underline{\theta},\, \underline{y})$ and $\underline{x} = \underline{\lambda}$; (iii) The classical evaluation of Methods of Moments estimators for which $\underline{\delta} = \underline{\theta}$ and $\underline{x} = (\underline{y},\, \underline{\lambda})$.

MC estimator of $G(\underline{\delta})$ in formula 13.15 is given by

$$\bar{G}_{S;p}(\underline{\delta}) = \frac{1}{S}\sum_{i=1}^{S} g(\underline{\tilde{x}}_i, \underline{\delta}) \ , \tag{13.16}$$

where the $\underline{\tilde{x}}$'s are i.i.d. draws from $p(\underline{x} \mid \underline{\delta})$ and S denotes the number of draws. The replacement of p by an alternative simulator with density μ calls for the following reformulation of $G(\underline{\delta})$

$$G(\underline{\delta}) = \int g(\underline{x}, \underline{\delta}) \cdot \omega(\underline{x}, \underline{\delta}) \cdot \mu(\underline{x} \mid \underline{\delta}) \, \mathrm{d}\underline{x} \, , \tag{13.17}$$

where

$$\omega(\underline{x}, \underline{\delta}) = \frac{p(\underline{x} \mid \underline{\delta})}{\mu(\underline{x} \mid \underline{\delta})} \, . \tag{13.18}$$

Note that the expectation of $\omega(\underline{X}, \underline{\delta})$ on μ equals one. The corresponding MC estimate of $G(\underline{\delta})$, known as an "importance sampling" estimate, is given by

$$\bar{G}_{S;\mu}(\underline{\delta}) = \frac{1}{S}\sum_{i=1}^{S} \omega(\underline{\tilde{x}}_i, \underline{\delta}) \cdot g(\underline{\tilde{x}}_i, \underline{\delta}) \ , \tag{13.19}$$

where the $\underline{\tilde{x}}_i$'s now are i.i.d. draws from $\mu(\underline{x} \mid \underline{\delta})$. The MC sampling variance of $\bar{G}_{S;\mu}(\underline{\delta})$ as an estimate of $G(\underline{\delta})$ is given by

$$V\left[\bar{G}_{S;\mu}(\underline{\delta})\right] = \frac{1}{S}\left\{\mathrm{E}_X\left[g^2(\underline{X}, \underline{\delta}) \cdot \omega^2(\underline{X}, \underline{\delta}) \mid \underline{\delta}\right] - G^2(\underline{\delta})\right\} \cdot \tag{13.20}$$

$\mu(\underline{x} \mid \underline{\delta})$ is the actual importance sampling density used in the construction of the EIS–MC estimates of the relevant integrals. (As we discuss in Sect. 13.4.2 below, p is instrumental in the construction of μ).

We will assume here that p is a genuine density function in that it integrates to one on its support $S(\underline{\delta})$. That assumption is routinely satisfied for classical inference procedures where $p(\underline{x} \mid \underline{\delta})$ represents a sampling distribution. In contrast it is frequently violated in Bayesian applications where $p(\underline{x} \mid \underline{\delta})$ represents a posterior density obtained by application of Bayes theorem which, therefore, often takes the form of a density kernel whose integrating constant is unknown and is itself to be numerically evaluated. Most quantities of interest are then ratios of integrals with the integral of p itself in the denominator.

In most applications $p(\underline{x} \mid \underline{\delta})$ will be a direct byproduct of the stochastic (sequential) specification of the model under consideration. We shall refer to it as an "initial sampler". For example, in Dynamic Latent Variables models, p typically corresponds to the marginal density of the latent process. See Hendry and Richard (1992) for details. Another example of initial sampler will be presented in sect. 13.6 below. Note that formulae (13.19) and (13.20) cover as special case that of the initial sampler with the simplification that if $\mu \equiv p$, then $\omega \equiv 1$.

Theoretically, an optimal choice μ_* for μ would be one such that the product $g \cdot \omega$ in formula (13.17) does not depend on $\underline{x}$. That is to say, if there exists an operational sampler $\mu_*(\underline{x} \mid \underline{\delta})$ and a "remainder" function $g_*(\underline{\delta})$ such that

$$g(\underline{x}, \underline{\delta}) \cdot p(\underline{x} \mid \underline{\delta}) = \mu_*(\underline{x} \mid \underline{\delta}) \cdot g_*(\underline{\delta}) \tag{13.21}$$

then $G(\underline{\delta}) \equiv g_*(\underline{\delta})$ and, furthermore, $V\left[\bar{G}_{S;\mu_*}(\underline{\delta})\right] = 0$ in which case a single draw from μ_* would produce the exact result. Note that when formula (13.21) holds, μ_* corresponds to the posterior density of $\underline{x}$ given $\underline{\delta}$. Except for the simplest models μ_* is generally not amenable to MC simulations. Furthermore, in high-dimensional problems, it is often the case that the MC sampling variance $V\left[\bar{G}_{S;p}(\underline{\delta})\right]$ is so large that accurate MC estimation of $G(\underline{\delta})$ using draws from the initial sampler p would require prohibitively large numbers of draws. See, in particular, the comments in McFadden (1989) as to the impracticability of simulated ML estimation in the context of discrete response models, or Danielsson and Richard (1993) for a striking example of the inaccuracy of naive MC estimates in the context of a stochastic volatility model. A number of "acceleration" techniques are available whereby the numerical accuracy of MC estimates can be enhanced, often at negligible increases in the cost of computations. See, e.g., Hendry (1984), Geweke (1988, 1994) or Davidson and McKinnon (1992). Nevertheless, these techniques constitute at best a partial remedy to the initial selection of an "inefficient" MC sampler and, under most circumstances, the only solution consists in the replacement of the initial sampler p by a more efficient importance sampler μ, i.e. one which is such that $G(\underline{\delta})$ can be accurately estimated with manageable number of draws. The literature on importance sampling provides useful examples of efficient samplers for specific classes of models. See Liesenfeld and Richard (2003a) for references. Until recently, however, there did not exist a generic algorithm to construct efficient importance samplers for (very) high-dimensional integrals of the type associated with high–frequency dynamic latent variable models (such as stochastic volatility models) and/or large panel models. Generalizing earlier results obtained by Danielsson and Richard (1993), Richard and Zhang (2007) proposed a generic least squares algorithm for the automated construction of Efficient Importance Samplers (EIS).

13.3.2 Efficient Importance Sampling

In this section we outline the general principle underlying EIS, refering the reader to Richard and Zhang (2007) for additional details. The specific implementation of EIS to panel data with unobserved heterogeneity will be presented in Sect. (13.6) below. We now assume that the function $g(\underline{x}, \underline{\delta})$ in (13.15) is strictly positive on the support $S(\underline{\delta})$, which, for example, is the case in all applications where $G(\underline{\delta})$ represents a likelihood function marginalized w.r.t. latent variables.

The construction of an EIS starts with the selection of a parametric class of samplers. Let $M = \{m(\underline{x} \mid \underline{\alpha});\ \underline{\alpha} \in A\}$ denote such a class. Typically, M would include parametric extensions of the initial sampler p designed to provide flexible approximations to the product $g \cdot p$ in (13.15), that is to say better approximations than p itself to the implicit posterior density $\mu_*(\underline{x} \mid \underline{\delta})$ in (13.21). Following (13.20), the selection of an "optimal" sampler within M is tantamount to solving the following optimization problem:

$$\underline{\alpha}^*(\underline{\delta}) = \arg\min_{\underline{\alpha} \in A} [V(\underline{\alpha};\ \underline{\delta})], \quad \text{with} \tag{13.22}$$

$$V(\underline{\alpha};\ \underline{\delta}) = \int \frac{\phi^2(\underline{x};\ \underline{\delta})}{m(\underline{x} \mid \underline{\alpha})} \mathrm{d}\underline{x} - G^2(\underline{\delta}) \tag{13.23}$$

$$\phi(\underline{x};\ \underline{\delta}) = g(\underline{x},\ \underline{\delta}) \cdot p(\underline{x} \mid \underline{\delta}) \tag{13.24}$$

The variance $V(\underline{\alpha};\ \underline{\delta})$ may be rewritten as

$$V(\underline{\alpha},\ \underline{\delta}) = G(\underline{\delta}) \cdot \int h\left[d^2(\underline{x};\ \underline{\delta},\ \underline{\alpha})\right] \cdot \phi(\underline{x};\ \underline{\delta})\, \mathrm{d}\underline{x} \tag{13.25}$$

with

$$d(\underline{x};\ \underline{\delta},\ \underline{\alpha}) = \ln\left[\frac{\phi(\underline{x};\ \underline{\delta})}{G(\underline{\delta}) \cdot m(\underline{x} \mid \underline{\alpha})}\right] \tag{13.26}$$

$$h(c) = e^{\sqrt{c}} + e^{-\sqrt{c}} - 2\,. \tag{13.27}$$

This EIS optimization problem can be significantly simplified further. First, we can replace $m(\underline{x} \mid \underline{\alpha})$ in (13.26) by a density kernel $k(\underline{x};\ \underline{\alpha})$. Let $\chi(\underline{\alpha})$ denote the integrating constant of that kernel, so that

$$m(\underline{x} \mid \underline{\alpha}) = \chi^{-1}(\underline{\alpha}) \cdot k(\underline{x};\ \underline{\alpha})\ . \tag{13.28}$$

Without loss of generality we then rewrite $d(\underline{x};\ \underline{\delta}, \underline{\alpha})$ as follows

$$d(\underline{x};\ \underline{\delta},\ \underline{\alpha}) = [\ln \phi(\underline{x};\ \underline{\delta}) - \gamma - \ln k(\underline{x};\ \underline{\alpha})]\ , \tag{13.29}$$

where $\gamma = \ln[G(\underline{\delta})/\chi(\underline{\alpha})]$ does not depend on $\underline{x}$ and is treated as an (additional) intercept in the optimization problem. Next, we note that if $m(\underline{x} \mid \underline{\alpha})$ belongs to the exponential family of distributions, then there exists an auxiliary reparametrisation such that $\ln k(\underline{x};\ \underline{\alpha})$ is linear in $\underline{\alpha}$, say

$$\ln k(\underline{x};\ \underline{\alpha}) = c(\underline{x}) \cdot \underline{\alpha}\,. \tag{13.30}$$

Finally, an efficient sampler will obviously be one such that $k(\underline{x};\ \underline{\alpha})$ closely mimics $\phi(\underline{x};\ \underline{\delta})$, in which $d(\underline{x};\ \underline{\delta},\ \underline{\alpha})$ is expected to be close to zero on average. Heuristically, this justifies replacing $h(c)$ in (13.27) by c, its leading term in a Taylor series expansion around zero. This amounts to approximating $V(\underline{\alpha};\ \underline{\delta})$ in (13.22) by

$$Q(\underline{\alpha};\,\underline{\delta}) = \int d^2(\underline{x};\,\underline{\delta},\,\underline{\alpha}) \cdot g(\underline{x},\,\underline{\delta}) \cdot p(\underline{x} \mid \underline{\delta})\,\mathrm{d}\underline{x} \tag{13.31}$$

resulting in a simpler Generalized Least Squares (GLS) optimization problem. Let $\hat{\underline{\alpha}}(\underline{\delta})$ denote the GLS solution to that problem. A more formal justification for the replacement of V by Q can be found in Richard and Zhang (2007) and follows from the inequality

$$V(\hat{\underline{\alpha}}(\underline{\delta});\,\underline{\delta}) \geq V(\underline{\alpha}_*(\underline{\delta});\,\underline{\delta}) \geq h\left[Q(\hat{\underline{\alpha}}(\underline{\delta});\underline{\delta})\right]. \tag{13.32}$$

An operational EIS implementation consists of solving first the simpler GLS problem and computing the two extreme bounds in (13.32) in order to assess whether additional efficiency gains would justify solving the computationally more demanding optimization problem in (13.25). Among all applications of EIS we have run over the last few years, including the one discussed in Sect. 13.6 below, there has never been one where the computation of $\underline{\alpha}_*(\underline{\delta})$ would have been justified. In practice, both optimization problems will be approximated by their finite sample MC counterparts. In particular, the MC version of the GLS optimization problem is given by

$$\hat{\underline{\alpha}}_R(\underline{\delta}) = \arg\min_{\alpha\in A\gamma\in R} \sum_{j=1}^{R} \left[\ln\phi\left(\tilde{\underline{x}}_j;\,\underline{\delta}\right) - \gamma - \ln k\left(\tilde{\underline{x}}_j;\,\underline{\alpha}\right)\right]^2 g\left(\tilde{\underline{x}}_j,\,\underline{\delta}\right) \tag{13.33}$$

where $\left\{\tilde{\underline{x}}_j;\ j:1\rightarrow R\right\}$ denotes i.i.d. draws from the initial sampler $p(\underline{x} \mid \underline{\delta})$. Since, in general, the MC sampling variance of $g(\underline{x},\,\underline{\delta})$ on draws from p is expected to be very large (which is why EIS is needed!) it is advisable to delete $g(\underline{x};\,\underline{\delta})$ from (13.33) and solving instead the LS problem. Note that, as typical within a LS framework, high variance in the draws from p actually helps securing an accurate global solution in (13.33).

In high–dimensional problems, the global optimization problem in (13.33) needs to be replaced by a sequence of manageable lower dimensional optimization problems. Here, we just outline the principle of such factorizations, refering the reader to Richard and Zhang (2007) for details and to Sect. 13.6 for a specific implementation. In line with (13.12), we assume that $\phi(\underline{x};\,\underline{\delta})$ is factorized as

$$\phi(\underline{x};\,\underline{\delta}) = \prod_{i=1}^{p} \phi_i(\underline{x}_{(i)};\,\underline{\delta}) \tag{13.34}$$

with

$$\phi_i(\underline{x}_{(i)};\,\underline{\delta}) = g_i(\underline{x}_{(i)};\,\underline{\delta}) \cdot p_i(\underline{x}_i \mid \underline{x}_{(i-1)},\,\underline{\delta}) \tag{13.35}$$

and the p_i's defining a (sequential) initial sampler. The importance sampler $m(\underline{x} \mid \underline{a})$ is partitioned conformably into

$$m(\underline{x} \mid \underline{a}) = \prod_{i=1}^{p} m_i(\underline{x}_i \mid \underline{x}_{(i-1)};\,\underline{a}_i), \quad \underline{a}_i \in A_i. \tag{13.36}$$

Let $k_i(\underline{x}_{(i)};\ \underline{a}_i)$ denote a kernel for m_i, with $\chi_i(\underline{x}_{(i-1)};\ \underline{a}_i)$ denoting its integrating constant (with respect to $\underline{x}_i$ only), whence

$$m_i(\underline{x}_i \mid \underline{x}_{(i-1)},\ \underline{a}_i) = \frac{k_i(\underline{x}_{(i)};\ \underline{a}_i)}{\chi_i(\underline{x}_{(i-1)};\ \underline{a}_i)}\ , \tag{13.37}$$

$$\text{with} \qquad \chi_i(\underline{x}_{(i-1)};\ \underline{a}_i) = \int k_i(\underline{x}_{(i)};\ \underline{a}_i)\mathrm{d}\underline{x}_i.$$

The key issue is that we can't expect to be able to approximate $\phi_i(\underline{x}_{(i)};\ \underline{\delta})$, whose integral in $\underline{x}_i$, is not known, by $m(\underline{x}_i \mid \underline{x}_{(i-1)},\ \underline{a}_i)$ which by definition integrates to one. Instead we could try to approximate ϕ_i by a kernel $k_i(\underline{x}_{(i)};\ \underline{a}_i)$, subject to the sole restriction that k_i has to be analytically integrable with respect to $\underline{x}_i$, so that once we have selected $\hat{\underline{a}}_i$ we have an analytical expression for χ_i. Obviously, by doing so, χ_i is not accounted for in the $\underline{a}_i$ LS optimization problem but, since it only depends on $\underline{x}_{(i-1)}$ it can be transfered back into the $\underline{a}_{i-1}$ LS optimization problem. In other words, the step i optimization subproblem consists of approximating the product $\phi_i\chi_{i+1}$ by a kernel k_i, specifically

$$\begin{aligned}\underline{\hat{\alpha}}_{i,R}(\underline{\delta}) = \arg \min_{\alpha_i \in A_i \gamma_i \in R} \sum_{j=1}^{R} \Big[&\ln\Big[\phi_i(\underline{\tilde{x}}_{(i),j};\ \underline{\delta}) \cdot \chi_{i+1}(\underline{\tilde{x}}_{(i),j};\ \underline{\hat{\alpha}}_{i+1,R}(\underline{\delta}))\Big] \\ &-\gamma_i - \ln k_i(\underline{\tilde{x}}_{(i),j};\ \underline{\alpha}_i)\Big]^2 \end{aligned} \tag{13.38}$$

for $i : p \to 1$ (with $\chi_{p+1} \equiv 1$), where $\{(\tilde{x}_{1,j}, \ldots, \underline{\tilde{x}}_{p,j});\ j : 1 \to R\}$ denotes i.i.d. "trajectories" drawn sequentially from $\{p_i(\underline{x}_i \mid \underline{\tilde{x}}_{(i-1),j};\ \underline{\delta})\}$. An example of such a sequential EIS implementation in the context of panel data is presented in Sect. 13.6 below.

13.3.3 MC Sampling Variance of (E)IS Estimates

A frequent criticism raised against (E)IS is the possibility that the variance $V(\underline{\alpha};\ \underline{\delta})$ in formula (13.23) might not exist, typically because the approximating kernel $k(\underline{x};\ \underline{\alpha})$ has thinner tails than the integrand $\phi(\underline{x};\ \underline{\delta})$. This criticism calls for important qualifications. Foremost, it applies to *all* MC methods relying upon approximations of the integrand, including Gibbs and Metropolis-Hastings procedures. Actually, if the variance of an IS estimate is infinite under a class M of auxiliary samplers, so will be the variance of any other MC estimate relying upon the same class. While this concern is addressed in the theoretical MCMC literature (see, e.g. Theorem 7.15 in Robert and Casella (2004)), the empirical MCMC literature largely ignores the possibility that an MCMC estimate might be invalid as the result of inappropriate selection of its auxiliary samplers.

Actually, Richard and Zhang (2007) propose a powerful test of the existence of the variance of an EIS estimate. This test consists of estimating $V(\underline{\alpha};\ \underline{\delta})$ in

formula (13.23) under two alternative samplers. It is an immediate by-product of the (E)IS evaluation and, most importantly, does not require any additional draws. Traditional verifications of the existence of the variance are typically based upon a very large number of draws in the hope of generating a very low probability 'outlier', which would destabilize the IS estimate. Such tests based upon the detection of a very rare event are notoriously unreliable. It is worth noting that the above mentioned test can also be applied to any other method relying upon auxiliary samplers, allowing for a very effective and unified test of variance finiteness in MC estimation.

13.3.4 GHK Simulator

The GHK simulator, to which we referred in our introduction, also belongs to a class of importance samplers. It is specifically designed to numerically evaluate probabilities of rectangles within a multivariate probit framework. It relies upon a triangular decomposition of the covariance matrix to construct an importance sampler in the form of a sequence of (conditional) univariate truncated Gaussian samplers (see, e.g., Gourieroux and Monfort (1994) for details). GHK has been widely and successfully applied to the numerical evaluation of likelihood functions for multivariate probit models.

In contrast with the EIS method described above, GHK is not designed to handle high-dimensional integrals nor does it includes an auxiliary optimization step (the latter is by no means as critical as for EIS in view of the lower dimensionality). Actually, GHK and EIS serve complementary purposes and can usefully be combined together to evaluate the likelihood of multivariate probit models with unobserved heterogeneity across individual units (in addition to the other dimensions handled by GHK). Specifically, in the context of formula (13.38), GHK would be used to evaluate lower dimensional integrals (e.g. multivariate probits) embedded in the expression of $\phi_i(\cdot)$. That is to say GHK would be used to evaluate low-dimensional inner integrals, while EIS would apply to (much higher-dimensional outer integrals. Such an application goes beyond the objectives of the present chapter but belongs to our research agenda. Also we intend to analyze in the future possibilities of incorporating within GHK an EIS-optimization step to increase its numerical efficiency.

13.3.5 Common Random Numbers

There was an important reason for carrying along $\underline{\delta}$ as an argument in all expressions from formula (13.15) onward: most (classical) simulation based inference procedures, some of which are discussed in Sect. 13.4 below, require the evaluation of a *function* $G(\underline{\delta})$ to be maximized (minimized) in $\underline{\delta}$.

As we replace $G(\underline{\delta})$ by its functional estimate $\hat{G}_S(\underline{\delta})$ an issue of smoothness immediately arises. Independent MC estimation of $G(\underline{\delta})$ at neighboring values of $\underline{\delta}$ would result in excessive wiggling of $\hat{G}_S(\underline{\delta})$, even when the latter is accurately estimated by EIS. This point is well recognized in the literature on simulated estimators—see, e.g. McFadden (1989), Pakes and Pollard (1989); Gourieroux and Monfort (1994)—where "smoothness" constitutes a critical condition for the validation of the asymptotic properties of these estimators. We will only consider here the obvious numerical problems that excessive wiggling would create for the optimization of $\hat{G}_S(\underline{\delta})$. Though we might consider smoothing $\hat{G}_S(\underline{\theta})$ in the process of optimization, a more direct remedy is available which takes advantage of the idiosyncrasies of pseudorandom number generation. The idea simply is that of reusing the same uniform pseudorandom numbers for the estimation of $G(\underline{\delta})$ at different $\underline{\delta}$'s, a technique which is known as that of "Common Random Numbers" (hereafter CRN's). Let $\left\{\underline{\tilde{x}}_{ij}, i : 1 \rightarrow S\right\}$ denote the random draws of $\underline{X}$ used for the estimation of $\hat{G}_S(\underline{\delta}_j)$. They can be obtained by transformation of a *common* sequence $\{\underline{\tilde{u}}_i;\ i : 1 \rightarrow S\}$ of uniform pseudorandom numbers, i.e.

$$\underline{\tilde{x}}_{ij} = \xi\left(\underline{\tilde{u}}_i;\ \underline{\delta}_j\right) . \tag{13.39}$$

This procedure will induce high positive correlations between estimates of $G(\underline{\delta})$ at neighboring values of $\underline{\delta}$. It will often suffice to secure sufficient smoothness for numerical optimization to succeed. Numerical evaluation of the derivatives of $\hat{G}_S(\underline{\delta})$ often remains delicate. Our own experience suggests using a "simplex" optimization algorithm. If analytical derivatives are available for the integrand in formula (13.15), MC estimates of the derivatives of $G(\underline{\delta})$ should be evaluated alongside with that of $G(\underline{\delta})$ itself.

The application of formula (13.39) requires attention when rejection techniques are being used as the actual number of uniform pseudorandom draws required to produce $\underline{\tilde{x}}_{ij}$ might vary with j. The simplest scenario is that when $\underline{X}$ is obtained by transformation of a "standardized" random variable $\underline{X}_*$ whose distribution does not depend on $\underline{\delta}$, i.e. when

$$\underline{X} = \xi_1\left(\underline{X}_*;\ \underline{\delta}\right) \quad \text{and} \quad \underline{X}_* = \xi_2\left(\underline{U}\right) . \tag{13.40}$$

In such cases, we only have to generate (or regenerate at will) a *single* sequence $\{\underline{\tilde{x}}_i^*\}$ from which the $\underline{\tilde{x}}_{ij}$'s are obtained by means of the transformation ξ_1 for all relevant values of $\underline{\delta}$. If, on the other hand, $\underline{X}$ is obtained directly from $\underline{U}$ and rejection is involved, it becomes very difficult to effectively implement CRN's. The alternative is then to rely upon the inversion technique, as defined in (13.2). Since, however, the inversion technique is typically much more time consuming than the most performant rejection techniques, its use can significantly increase overall computing time. Taking full advantage of the typical very low dimensionality of sequential EIS kernels, careful reliance upon numerical interpolation techniques can usefully be considered – though such discussion goes beyond the objectives of this chapter.

13.4 Simulation-Based Inference Procedures

In this section we briefly discuss the use of MC integration in the context of three of the most commonly used inference techniques: Maximum likelihood, (generalized) Method of Moments and Bayesian posterior moments. Details of implementation are highly problem dependent and will not be considered here.

13.4.1 Integration in Panel Data Models

There exists a growing number of panel data applications requiring multidimensional integrations for which no analytical solutions are available and quadrature (non–stochastic) rules are impractical. A few examples are listed below:

(a) Multinomial probit models involve integrals whose dimensionality equals the number of alternatives minus one;
(b) The elimination of latent variables in nonlinear models also requires numerical integration. The dimensionality of integration typically is that of the vector of latent variables;
(c) Relatedly, the elimination of individual unobserved heterogeneity factors is also done by integration. Though it is occasionally possible to find analytical solutions under specific choices of distributions, numerical integration is required at a higher level of generality;
(d) The evaluation of a likelihood function for panel data with missing observations necessitates the computation of (multidimensional) integrals which, except for special cases, has to be done numerically.

Note also that the application of Bayesian inference techniques to any of these models generally requires additional (numerical) integrations with respect to the parameters themselves. Such an application is discussed in Sect. 25.3 in the context of Markov processes. Finally, there are numerous extensions of the models specifically discussed in this handbook that would require (additional) numerical integrations.

Simulation techniques are increasingly used to evaluate such integrals. A few references of interest in the context of panel data models are Lerman and Manski (1981), McFadden (1989), Pakes and Pollard (1989) or Börsch–Supan and Hajivassiliou (1993). These contributions and others have led to the development of a number of simulation based inference techniques some of which are briefly described below. See also Gourieroux and Monfort (1993) for a recent survey with reference to panel data or Gourieroux and Monfort (1994) for an in-depth analysis of simulation based econometric techniques.

Two key features characterize this line of work. Firstly, it often relies upon non "efficient" Monte-Carlo procedures (in the numerical sense). Some of the comments found in the literature as to the impracticability of MC likelihood evaluation – see, e.g. McFadden (1989) – have to be qualified in that context. Secondly, it conflates

"statistical" and "numerical" properties of simulation estimators, a point we will address further in Sect. 13.5 below.

13.4.2 Simulated Likelihood

Within a likelihood framework, $\underline{\delta}$ in formula (13.15) consists of observables $\underline{y}$ and unknown parameters $\underline{\theta}$, while $\underline{\lambda}$ regroups all unobservables. The simplest case is that where a "marginalized" likelihood has to be evaluated which is of the form

$$L\left(\underline{\theta};\, \underline{y}\right) = \int f\left(\underline{y},\, \underline{\lambda} \mid \underline{\theta}\right) \mathrm{d}\underline{\lambda}\,, \tag{13.41}$$

where f denotes the joint density of all relevant variables, observables *and* unobservables. In most cases f takes the form of a product of component densities, e.g. in the form of (13.12) or (13.13). Such factorizations may considerably simplify the actual implementation of a simulation algorithm but will not be specifically considered here. Let $\hat{L}_S(\underline{\theta};\, \underline{y})$ denote a functional MC estimator of $L(\underline{\theta};\underline{y})$ obtained by MC simulation. A simulated maximum likelihood (SML) estimator of θ is given by

$$\underline{\hat{\theta}}_S\left(\underline{y}\right) = \arg\max_{\theta} \left[\ln \hat{L}_S\left(\underline{\theta};\, \underline{y}\right)\right]\,. \tag{13.42}$$

An example of SML estimation using EIS with be presented in Sect. 13.6 below.

There are numerous important problems (such as discrete choice models) where the likelihood function itself is not in the form of an integral but depends upon integrals which have to be numerically evaluated. A general formulation of such problems would take the following form

$$L\left(\underline{\theta};\, \underline{y}\right) = h\left(\underline{G}\left(\underline{y} \mid \underline{\theta}\right);\, \underline{y},\, \underline{\theta}\right), \quad \text{with} \tag{13.43}$$

$$\underline{G}\left(\underline{y} \mid \underline{\theta}\right) = \int \underline{\gamma}\left(\underline{y},\, \underline{\lambda} \mid \underline{\theta}\right) \mathrm{d}\underline{\lambda}\,. \tag{13.44}$$

Examples can be found, e.g. in Gourieroux and Monfort (1994), together with extensions to simulated pseudo maximum likelihood estimation.

13.4.3 Simulated Method of Moments

There exist many excellent discussions of the Generalized Method of Moments (GMM) and of the Simulated Method of Moments (MSM) in the recent literature. The short presentation which follows draws upon Davidson and McKinnon (1993) for GMM and Pakes and Pollard (1989) for MSM. See also McFadden (1989) or Gourieroux and Monfort (1994). Let

$$\underline{G}(\underline{\theta}) = \int \underline{H}\left(\underline{y},\, \underline{\theta}\right) \cdot f_Y\left(\underline{y} \mid \underline{\theta}\right) \mathrm{d}\underline{y} \tag{13.45}$$

together with

$$\underline{H}\left(\underline{y},\, \underline{\theta}\right) = \int \underline{h}\left(\underline{y},\, \underline{\lambda};\, \underline{\theta}\right) \cdot f_\Lambda\left(\underline{\lambda} \mid \underline{y},\, \underline{\theta}\right) \mathrm{d}\underline{\lambda} \tag{13.46}$$

denote a set of moments conditions assumed to be zero at the true value $\underline{\theta}_0$. The empirical conditions associated with (13.45) are given by

$$\sum_{j=1}^{NT} \underline{H}(\underline{y}_j,\, \underline{\theta}_0) = 0\,, \tag{13.47}$$

where $\{\underline{y}_j;\ j : 1 \to NT\}$ denotes the actual sample. (The "short hand" notation used in (13.45) appears to suggest that the $\underline{y}_j$'s are i.i.d. with density $f_Y(\underline{y} \mid \underline{\theta})$. Results are also available for dependent y's. See, e.g., Gallant (1987). Let $Y' = (\underline{y}_1, \ldots, \underline{y}_{NT})$. A GMM estimator of θ is given by

$$\hat{\underline{\theta}}(Y) = \arg\min_{\underline{\theta}} \left[\sum_{j=1}^{NT} \underline{H}(\underline{y}_j,\, \underline{\theta})\right]' A(Y) \left[\sum_{j=1}^{NT} \underline{H}(\underline{y}_j,\, \underline{\theta})\right]\,, \tag{13.48}$$

where $A(Y)$ is a symmetric positive definite matrix. The consistency of $\hat{\underline{\theta}}(Y)$ obtains under a broad range of conditions. Efficiency requires that

$$\operatorname*{plim}_{NT\to\infty} A(Y) = A\mathrm{Var}\left[\frac{1}{\sqrt{NT}} \sum_{j=1}^{NT} \underline{H}(\underline{y}_j,\, \underline{\theta}_0)\right] := A_0\,. \tag{13.49}$$

The asymptotic variance of $\sqrt{NT}\,\hat{\underline{\theta}}(Y)$ on condition (13.49) is given by

$$A\mathrm{Var}\left[\sqrt{NT}\,\hat{\underline{\theta}}\,(Y)\right] = \left(D_0 A_0^{-1} D_0'\right)^{-1}, \quad \text{with} \tag{13.50}$$

$$D_0 = \operatorname*{plim}_{NT\to\infty} \left[\frac{1}{NT} \sum_{j=1}^{NT} \frac{\partial \underline{H}'(\underline{y}_j,\, \underline{\theta})}{\partial \underline{\theta}}\right]_{\underline{\theta}=\underline{\theta}_0}\,. \tag{13.51}$$

Let $\underline{\hat{H}}_S\,(\underline{y}_j,\, \underline{\theta})$ denote a (convergent) MC estimator of $\underline{H}(\underline{y},\, \underline{\theta})$. An MSM estimator of $\underline{\theta}$ is given by

$$\hat{\underline{\theta}}_S(Y) = \arg\min_{\underline{\theta}} \left[\sum_{j=1}^{NT} \underline{\hat{H}}_S(\underline{y}_j,\, \underline{\theta})\right]' A(Y) \left[\sum_{j=1}^{NT} \underline{\hat{H}}_S(\underline{y}_j;\, \underline{\theta})\right]\,. \tag{13.52}$$

As discussed earlier it is important that CRN's be used in drawing the $\underline{\lambda}$'s from $f_\Lambda(\underline{\lambda} \mid \underline{y},\, \underline{\theta})$.

13.4.4 Bayesian Posterior Moments

Let $\nu(\underline{\theta})$ denote a prior density for the parameters of the marginalized likelihood $L(\underline{\theta};\ \underline{y})$, as definded in (13.41). The posterior density of $\underline{\theta}$ is proportional to the product of ν by L and its integrating constant is generally unknown. The posterior density of a function $h(\underline{\theta})$ is given by

$$h^* = \frac{\int\int h(\underline{\theta})\, f\left(\underline{y},\ \underline{\lambda} \mid \underline{\theta}\right) \nu(\underline{\theta})\, \mathrm{d}\underline{\lambda}\mathrm{d}\underline{\theta}}{\int\int f\left(\underline{y},\ \underline{\lambda} \mid \underline{\theta}\right) \nu(\underline{\theta})\, \mathrm{d}\underline{\lambda}\mathrm{d}\underline{\theta}}\ . \tag{13.53}$$

A convergent (E)IS estimator of h^* is given by

$$\hat{h}_S^* = \frac{\sum_{i=1}^{S} \omega(\tilde{\underline{\lambda}}_i,\ \tilde{\underline{\theta}}_i,\ \underline{y}) h(\tilde{\underline{\theta}}_i)}{\sum_{i=1}^{S} \omega(\tilde{\underline{\lambda}}_i,\ \tilde{\underline{\theta}}_i;\ \underline{y})}\ , \tag{13.54}$$

where $\{(\tilde{\underline{\lambda}}_i,\ \tilde{\underline{\theta}}_i);\ i : 1 \to S\}$ are i.i.d. draws from a sampler $\mu(\underline{\lambda},\ \underline{\theta} \mid \underline{y})$ and $\omega = f\nu/\mu$. Details are found e.g., in Geweke (1994).

Note that the evaluation of $\hat{h}_S^*$ only requires a single set of *joint* draws from μ. In contrast, maximization of $\hat{L}_S(\underline{\theta};\underline{y})$ in (13.42) requires a new set of (CRN) draws of $\underline{\lambda}$ for each value of $\underline{\theta}$ at which the likelihood has to be evaluated.

13.5 Numerical Properties of Simulated Estimators

There is a fundamental difference between the ways in which classical and Bayesian econometricians evaluate the properties of simulated estimators. Bayesians treat the actual sample $\underline{y}$ as fixed. The only source of randomness to be accounted for originates from the auxiliary MC sample. It follows, in particular, that Bayesians routinely assess the *numerical* accuracy of their MC estimates of quantities of interest, e.g. in the form of MC standard deviations.

In contrast classical econometricians insist upon treating $\hat{\underline{\theta}}_S(\underline{y})$ as an estimate of $\underline{\theta}$ itself and, therefore, have to account for two independent sources of randomness: the actual sample $\underline{y}$ and the auxiliary MC sample $\underline{\lambda}$. We find three major drawbacks to the classical analysis of simulated estimators:

(i) It confuses the issue of assessing the *statistical* properties of $\hat{\underline{\theta}}(y)$ as an estimate of $\underline{\theta}$ with that of evaluating the *numerical* accuracy of $\hat{\underline{\theta}}_S(y)$ as an MC estimate of $\hat{\underline{\theta}}(y)$;

(ii) It complicates the analysis of simulation estimators since the two sources of randomness apply at fundamentally different levels. Specifically, the statistical properties of $\hat{\underline{\theta}}(y)$ are determined by data availability. In contrast, the numerical properties of $\hat{\underline{\theta}}_S(y)$ are fully controlled by the analyst and can be arbitrarily

reduced either by increasing MC sample sizes and/or, more efficiently, by increasing the efficiency of the simulation techniques (which is precisely what EIS has been designed for).

(iii) It applies to situations where the observables y and the latent variables λ are *jointly* simulated. However, as we argued above, integration of λ can often be carried out much more efficiently by using EIS in which case numerical accuracy is typically far greater than statistical accuracy and conventional formulae do not apply.

Nowhere is the problem more apparent than in the context of simulated ML estimators. The statistical properties of an ML estimator $\hat{\underline{\theta}}(\underline{y})$ are well understood and follow from the application of a Central Limit Theorem (hereafter CLT) to the derivatives of the *logarithm* of the likelihood function. In contrast MC estimation applies to the likelihood function itself. Since integrals and logarithms do not commute, it follows that $\hat{\underline{\theta}}_S(\underline{y})$ is an inconsistent estimator of $\underline{\theta}$ for fixed S and comments to that effect abound in the literature.

We propose instead to keep treating $\hat{\underline{\theta}}(\underline{y})$ as a statistical estimator of $\theta(y)$ and, when $\hat{\underline{\theta}}(\underline{y})$ cannot be computed, to treat $\hat{\underline{\theta}}_S(\underline{y})$ as a numerical estimator of $\hat{\underline{\theta}}(\underline{y})$, *not* of $\underline{\theta}$. Statistical inference then follows standard procedures. For example, we already know that, under appropriate conditions,

$$\sqrt{NT}\left[\hat{\underline{\theta}}\left(\underline{y}\right)-\underline{\theta}\right] \xrightarrow{d} N\left(\underline{0}, V\left(\underline{\theta}\right)\right) , \tag{13.55}$$

where NT is actual sample size and $V(\underline{\theta})$ is a covariance matrix to be estimated by a matrix $\hat{V}(\underline{y})$. When no analytical expressions are available for the likelihood function, we can use MC simulation as a *numerical* device for computing approximations to $\hat{\underline{\theta}}(\underline{y})$ and $\hat{V}(\underline{y})$. At this stage of the analysis, our sole problem is that of assessing the *numerical accuracy* of these approximations which are to be treated as functions of $\underline{\lambda}$, *given* $\underline{y}$.

For large S's, we can apply standard techniques to obtain "asymptotic" MC sampling distributions for $\hat{\underline{\theta}}_S(\underline{y})$—here again $\underline{y}$ is kept fixed at its observed value. Under conditions such as those found, e.g., in Geweke (1994), a CLT applies to $\hat{L}_s(\underline{\theta};\, \underline{y})$—not to its logarithm—and to its derivatives. In particular,

$$\sqrt{S}\left[\frac{\partial \hat{L}_S\left(\underline{\theta};\, \underline{y}\right)}{\partial \underline{\theta}}-\frac{\partial L\left(\underline{\theta};\, \underline{y}\right)}{\partial \underline{\theta}}\right] \xrightarrow{d} N\left(\underline{0}, \Omega\left(\underline{\theta};\, \underline{y}\right)\right) , \tag{13.56}$$

where $\Omega(\underline{\theta};\, \underline{y})$ is a covariance matrix that can be estimated alongside with $\hat{\underline{\theta}}_S(\underline{y})$. It follows that

$$\sqrt{S}\left[\hat{\underline{\theta}}_S\left(\underline{y}\right)-\hat{\underline{\theta}}\left(\underline{y}\right)\right] \xrightarrow{d} N\left(\underline{0}, P^{-1}\left(\underline{\theta};\, \underline{y}\right)\Omega\left(\underline{\theta};\, \underline{y}\right)P^{-1}\left(\underline{\theta};\, \underline{y}\right)'\right) \tag{13.57}$$

with

$$P\left(\underline{\theta};\, \underline{y}\right)=\operatorname*{plim}_{S\to\infty}\frac{1}{S}\frac{\partial^2 \hat{L}_S\left(\underline{\theta};\, \underline{y}\right)}{\partial \underline{\theta}\partial \underline{\theta}'} . \tag{13.58}$$

Estimates of $P(\underline{\theta};\ \underline{y})$ and $\Omega(\underline{\theta};\ \underline{y})$ can be obtained as byproducts of the simulation runs.

In general, however, it is much easier to produce "finite sample" numerical covariance matrices. All that is required is to rerun the entire MC–ML algorithm under different seeds for the $\underline{\lambda}$'s, producing thereby i.i.d. draws of $\hat{\underline{\theta}}_S(\underline{y})$ for a given $\underline{y}$. Similarly finite sample statistical covariance matrices can be obtained by simulation of $\underline{y}$ for a *given set* of $\underline{\lambda}$ CRNs. (That is to say, the $\underline{\lambda}$ draws used for any particular $\underline{y}$ and which, for reasons of numerical efficiency need to be conditional upon $\underline{y}$, are to be obtained by transformation of a *fixed* set $\underline{u}$ of canonical draws—see (13.39).)

In other words, once a MC simulation program has been produced to compute simulated ML estimators for a specific problem, it is generally trivial to produce separate estimates of numerical and statistical accuracy by embedding that program into two distinct external simulation loops (one for $\underline{\lambda}(\underline{U})$ given $\underline{y}$ and the other for $\underline{y}$ given $\underline{U}$). Note that it is also possible to construct joint simulation of $\underline{\lambda}(\underline{U})$ and $\underline{y}$ to produce a measure of the compound uncertainty of $\hat{\underline{\theta}}_S(\underline{y})$ as an estimate of $\underline{\theta}$.

13.6 EIS Application: Logit Panel with Unobserved Heterogeneity

13.6.1 The Model

In the following, we discuss the application of EIS to the evaluation of the likelihood of a panel logit model with unobserved heterogeneity, illustrating the full sequential implementation of the procedure. Other applications of EIS are, for example, the estimation of various univariate and multivariate specifications of the stochastic volatility model for financial returns by Liesenfeld and Richard (2003c), the estimation of dynamic parameter-driven count-data models by Jung and Liesenfeld (2001) and the estimation of stochastic autoregressive intensity processes for financial market activities on a trade-by-trade basis by Bauwens and Hautsch (2003).

Consider the following model for the latent variable y^*_{ti} for individual i and time period t:

$$y^*_{ti} = \underline{\beta}' \underline{z}_{ti} + \tau_i + \lambda_t + \varepsilon_{ti}\,, \qquad i : 1 \to N, \quad t : 1 \to T\,, \tag{13.59}$$

where $\underline{z}_{ti}$ is a vector of explanatory variables and ε_{ti} is an i.i.d. logistic variable with zero mean and variance $\pi^2/3$. τ_i represents individual random effects and is assumed to be i.i.d. Gaussian with zero mean and variance σ^2_τ. λ_t captures time random effects and is assumed to follow a stationary autoregressive process

$$\lambda_t = \rho_1 \lambda_{t-1} + \cdots + \rho_k \lambda_{t-k} + \eta_t\,, \tag{13.60}$$

where η_t is an i.i.d. Gaussian variable with zero mean and variance σ^2_η such that the stationary mean of λ_t is zero and the stationary variance is $\sigma^2_\lambda = \sigma^2_\eta/(1-\rho_1^2-\cdots-\rho_k^2)$. It is assumed that the components ε_{ti}, τ_i, and η_t are mutually independent. The

observable variable is given by the dummy variable $y_{ti} = I_{\{y_{ti}^* \geq 0\}}$ and the vector of parameters to be estimated is $\underline{\theta} = (\beta', \sigma_\tau, \rho_1, \ldots, \rho_k, \sigma_\lambda)'$. Let $\underline{\lambda} = (\lambda_1, \ldots, \lambda_T)'$ and $\underline{\tau} = (\tau_1, \ldots, \tau_N)'$, then the likelihood function associated with $\underline{y} = (y_{11}, \ldots, y_{TN})'$ can be written as

$$L(\underline{\theta};\, \underline{y}) = \int g(\underline{\tau}, \underline{\lambda};\, \underline{\theta}, \underline{y})\, p(\underline{\tau}, \underline{\lambda};\, \underline{\theta})\, \mathrm{d}\underline{\tau}\mathrm{d}\underline{\lambda}\,, \tag{13.61}$$

with

$$g(\underline{\tau}, \underline{\lambda};\, \underline{\theta}, \underline{y}) = \prod_{i=1}^{N}\prod_{t=1}^{T} \left[\frac{1}{1+\exp\{v_{ti}\}}\right]^{1-y_{ti}} \left[\frac{\exp\{v_{ti}\}}{1+\exp\{v_{ti}\}}\right]^{y_{ti}} \tag{13.62}$$

$$p(\underline{\tau}, \underline{\lambda};\, \underline{\theta}) \propto \sigma_\tau^{-N} \exp\left[-\frac{1}{2\sigma_\tau^2}\sum_{i=1}^{N}\tau_i^2\right] |\Sigma_\lambda|^{-1/2} \cdot \exp\left[-\frac{1}{2}\underline{\lambda}'\Sigma_\lambda^{-1}\underline{\lambda}\right] \tag{13.63}$$

$$v_{ti} = \underline{\beta}'\underline{z}_{ti} + \tau_i + \lambda_t\,, \tag{13.64}$$

where Σ_λ denotes the stationary variance–covariance matrix of $\underline{\lambda}$.

A natural MC estimator of this likelihood function for given values of $\underline{\theta}$ and $\underline{y}$ is given by

$$\bar{L}_{S;p}(\underline{\theta};\, \underline{y}) = \frac{1}{S}\sum_{r=1}^{S} g(\underline{\tilde{\tau}}_r, \underline{\tilde{\lambda}}_r; \underline{\theta}, \underline{y})\,, \tag{13.65}$$

where $\{(\underline{\tilde{\tau}}_r, \underline{\tilde{\lambda}}_r);\ r : 1 \to S\}$ are i.i.d. draws from $p(\underline{\tau}, \underline{\lambda};\, \underline{\theta})$. Since the natural sampling density p directly obtained from the statistical formulation of the model, does not incorporate critical information about the latent processes conveyed by the observations $\underline{y}$, the natural estimator $\bar{L}_{S;p}$ is highly inefficient. In fact, for all practical purposes, a prohibitively large MC sample size would be required to obtain reasonably accurate estimates of L. Moreover, the implicit "posterior" density of $(\underline{\tau},\ \underline{\lambda})$ is much tighter than its "prior" (the natural sampler), since the sample conveys significant information on unobserved heterogeneity. Whence the "important" domain of integration where g effectively contributes to the value of the integral is much tighter than that implied by the natural sampler. Thus, the probability that a MC draw $(\underline{\tilde{\tau}}_r, \underline{\tilde{\lambda}}_r)$ hits exactly this region is almost zero and, hence, the natural MC estimate $\bar{L}_{S;p}$ is severely downward biased.

13.6.2 EIS Evaluation of the Likelihood

As discussed above (see formulae (13.34) to (13.37)), the global high–dimensional EIS optimization problem (13.33) associated with $L(\underline{\theta};\, \underline{y})$ has to be broken down

into low-dimensional subproblems according to an appropriate factorization of the integrand $\phi(\underline{\tau},\underline{\lambda};\ \underline{\theta},\underline{y}) = g(\underline{\tau},\underline{\lambda};\ \underline{\theta},\underline{y})p(\underline{\tau},\underline{\lambda};\ \underline{\theta})$. Since the empirical application discussed below is based on a data set with $N \gg T$, the integrand will be factorized into a product of N terms each of which depends upon a single τ_i and a remainder (for $T \gg N$, one would instead factorize ϕ into T terms, each of which depends upon a single λ_t)

$$L\left(\underline{\theta};\ \underline{y}\right) = \int \phi_0\left(\underline{\lambda};\ \underline{\theta}\right) \prod_{i=1}^{N} \phi_i\left(\tau_i,\underline{\lambda};\ \underline{\theta},\underline{y}\right) \mathrm{d}\underline{\tau}\mathrm{d}\underline{\lambda}\,, \tag{13.66}$$

with

$$\phi_i\left(\tau_i,\underline{\lambda};\ \underline{\theta},\underline{y}\right) \propto \sigma_\tau^{-1} \exp\left[-\frac{\tau_i^2}{2\sigma_\tau^2}\right] \prod_{t=1}^{T} \left[\frac{1}{1+\exp\{v_{ti}\}}\right]^{1-y_{ti}} \cdot \left[\frac{\exp\{v_{ti}\}}{1+\exp\{v_{ti}\}}\right]^{y_{ti}}, \tag{13.67}$$

$$\phi_0(\underline{\lambda};\ \underline{\theta}) \propto |\Sigma_\lambda|^{-1/2} \exp\left[-\frac{1}{2}\underline{\lambda}'\Sigma_\lambda^{-1}\underline{\lambda}\right]\,. \tag{13.68}$$

According to this factorization, the global optimization problem associated with ϕ can be factorized into subproblems each of which is related to a single component ϕ_i $(i: 0 \to N)$ whose salient characteristics ought to be captured by a corresponding efficient sampling density. Note that, even though the τ_i's and λ_t's are, according to the assumptions of the statistical model, stochastically independent, ϕ_i introduces interdependencies between these variables. In order to take this into account the efficient sampler can be constructed as a sequence of sampling densities with an unconditional density for $\underline{\lambda}$ and a sequence of conditional densities for $\tau_i \mid \underline{\lambda}$. The resulting factorization of the efficient sampler is given by

$$m(\underline{\tau},\underline{\lambda} \mid \underline{\alpha}) = m_0(\underline{\lambda};\ \underline{\alpha}_0) \prod_{i=1}^{N} m_i(\tau_i \mid \underline{\lambda};\ \underline{\alpha}_i)\,, \tag{13.69}$$

where m_0 and $\{m_i\}$ are specified (as a natural choice) as parametric extensions of the corresponding densities defining the natural sampler p, and $\underline{\alpha} = (\underline{\alpha}_0', \underline{\alpha}_1', \ldots, \underline{\alpha}_N')'$ is a vector of auxiliary parameters. For any given value of $\underline{\alpha}$, the likelihood (13.66) can be rewritten as

$$L\left(\underline{\theta};\ \underline{y}\right) = \int \frac{\phi_0\left(\underline{\lambda};\ \underline{\theta}\right)}{m_0\left(\underline{\lambda};\ \underline{\alpha}_0\right)} \prod_{i=1}^{N} \left[\frac{\phi_i\left(\tau_i,\underline{\lambda};\ \underline{\theta},\underline{y}\right)}{m_i\left(\tau_i \mid \underline{\lambda};\ \underline{\alpha}_i\right)}\right] m(\underline{\tau},\underline{\lambda} \mid \underline{\alpha})\,\mathrm{d}\underline{\tau}\mathrm{d}\underline{\lambda}\,, \tag{13.70}$$

and the corresponding MC estimate of the likelihood is given by

$$\tilde{L}_{S;m}\left(\underline{\theta};\ \underline{y},\underline{\alpha}\right) = \tag{13.71}$$

$$\frac{1}{S}\sum_{r=1}^{S} \frac{\phi_0(\tilde{\underline{\lambda}}_r(\underline{\alpha}_0);\ \underline{\theta})}{m_0(\tilde{\underline{\lambda}}_r(\underline{\alpha}_0);\ \underline{\alpha}_0)} \prod_{i=1}^{N} \frac{\phi_i(\tilde{\tau}_{ir}(\alpha_i), \tilde{\underline{\lambda}}_r(\underline{\alpha}_0);\ \underline{\theta},\underline{y})}{m_i(\tilde{\tau}_{ir}(\alpha_i) \mid \tilde{\underline{\lambda}}_r(\underline{\alpha}_0);\ \underline{\alpha}_i)}\,,$$

where $\{[\tilde{\tau}_{1r}(\alpha_1),\ldots,\tilde{\tau}_{Nr}(\alpha_N),\tilde{\underline{\lambda}}_r(\underline{\alpha}_0)];\ r:1\rightarrow S\}$ are i.i.d. draws from the auxiliary importance sampling density $m(\underline{\tau},\underline{\lambda}|\underline{\alpha})$.

As discussed above, the optimal sampling density m_i requires constructing a functional approximation $k_i(\tau_i;\ \underline{\lambda},\underline{\alpha}_i)$ for $\phi_i(\tau_i,\underline{\lambda};\ \underline{\theta},\underline{y})$ with the requirement that its integral with respect to τ_i (depending upon $\underline{\lambda}$) can be computed analytically. Specifically, the function $k_i(\tau_i;\ \underline{\lambda},\underline{\alpha}_i)$ serves as a density kernel for $m_i(\tau_i|\underline{\lambda};\ \underline{\alpha}_i)$ which is given by

$$m_i(\tau_i\mid\underline{\lambda};\ \underline{\alpha}_i)=\frac{k_i(\tau_i;\ \underline{\lambda},\underline{\alpha}_i)}{\chi_i(\underline{\lambda},\underline{\alpha}_i)},\quad\text{where}\quad\chi_i(\underline{\lambda},\underline{\alpha}_i)=\int k_i(\tau_i;\ \underline{\lambda},\underline{\alpha}_i)\,\mathrm{d}\tau_i\,. \tag{13.72}$$

Note that a good match between the ϕ_i's alone and the k_i's would leave the sequence of χ_i's unaccounted for. But since the χ_i's do not depend on the τ_i's they can be attached to the problem of matching ϕ_0 by m_0. Accordingly, the likelihood can be rewritten as

$$L(\underline{\theta};\ \underline{y})=\int\frac{\phi_0(\underline{\lambda};\ \underline{\theta})\prod_{i=1}^N\chi_i(\underline{\lambda};\ ,\underline{\alpha}_i)}{m_0(\underline{\lambda};\ \underline{\alpha}_0)}\prod_{i=1}^N\left[\frac{\phi_i(\tau_i,\underline{\lambda};\ \underline{\theta},\underline{y})}{k_i(\tau_i;\ \underline{\lambda};\ \underline{\alpha}_i)}\right] \tag{13.73}$$

$$\cdot m(\underline{\tau},\underline{\lambda}\mid\underline{\alpha})\,\mathrm{d}\underline{\tau}\mathrm{d}\underline{\lambda}\,.$$

Taken all together, the sequential implementation of the global high–dimensional EIS optimization problem requires solving a sequence of $N+1$ low-dimensional (weighted) LS problems of the form

$$\begin{aligned}\hat{\underline{\alpha}}_i=\arg\min_{\underline{\alpha}_i}\sum_{r=1}^S\Big\{&\ln\phi_i(\tilde{\tau}_{ir},\tilde{\underline{\lambda}}_r;\ \underline{\theta},\underline{y})-c_i\\&-\ln k_i(\tilde{\tau}_{ir};\ \tilde{\underline{\lambda}}_r,\underline{\alpha}_i)\Big\}^2 g_i(\tilde{\tau}_{ir},\tilde{\underline{\lambda}}_r;\ \underline{\theta},\underline{y})\end{aligned} \tag{13.74}$$

for $i:1\rightarrow N$ and

$$\begin{aligned}\hat{\underline{\alpha}}_0=\arg\min_{\underline{\alpha}_0}\sum_{r=1}^S\Bigg\{&\ln[\phi_0(\tilde{\underline{\lambda}}_r;\ \underline{\theta})\prod_{i=1}^N\chi_i(\tilde{\underline{\lambda}}_r,\hat{\underline{\alpha}}_i)]-c_0\\&-\ln m_0(\tilde{\underline{\lambda}}_r;\ \underline{\alpha}_0)\Bigg\}^2,\end{aligned} \tag{13.75}$$

where the weights $\{g_i(\tau_i,\underline{\lambda};\ \underline{\theta},\underline{y});\ i:1\rightarrow N\}$ are given by the N terms of the outer product of the function $g(\underline{\tau},\underline{\lambda};\ \underline{\theta},\underline{y})$ given in (13.62) (the weight for the LS problem (13.75) is by construction $g_0(\cdot)=1$). $\{(\tilde{\tau}_{1r},\ldots,\tilde{\tau}_{Nr},\tilde{\underline{\lambda}}_r),\ r:1\rightarrow S\}$ are i.i.d. draws from the natural sampler p, and the c_i's and c_0 are unknown constants to be estimated jointly with the $\underline{\alpha}_i$'s and $\underline{\alpha}_0$.

One iteration of the EIS optimization algorithm generally suffices to produce a vastly improved importance sampler. Nevertheless, a small number of iterations where the natural sampler p for the (weighted) LS problems is replaced by the

previous stage importance sampler produces further efficiency gains. For such iterations to converge to a fixed value of auxiliary parameters $\underline{\alpha}$, which are expected to produce the optimal sampler, it is necessary to use CRNs implying that all draws for the τ_i's and $\underline{\lambda}$ for any given sampling density are obtained from a fixed set of standardized random numbers (see, Sect. (13.3.3)). Also, as discussed earlier, it is generally preferable to set the LS weights in (13.74) equal to one in the first iteration.

Furthermore, observe that for the logit model with Gaussian random effects the use of parametric extensions of p for the construction of the efficient sampler m implies that the k_i's are Gaussian density kernels for the τ_i's given $\underline{\lambda}$, and m_0 be a multivariate Gaussian density for $\underline{\lambda}$. Thus, the LS problems (13.74) and (13.75) become linear in $\underline{\alpha}_i$ and $\underline{\alpha}_0$. Moreover, note that in this case ϕ_0 as well as the χ_i's are Gaussian kernels for $\underline{\lambda}$ allowing for a perfect fit in the LS problem (13.75) and for an analytical construction of the efficient sampler m_0 for $\underline{\lambda}$. (For a full description of the implementation of the efficient sampling procedure for the logit panel model, see the Appendix.)

Finally, the MC likelihood estimate for any admissible value of $\underline{\theta}$ and $\underline{y}$ based on the optimal sampler is obtained by substituting $\{\underline{\alpha}_i\,, i:0\rightarrow N\}$ by $\{\underline{\hat{\alpha}}_i\,, i:0\rightarrow N\}$ in equation (13.71). Based on the efficient sampling procedure, the ML parameter estimates of $\underline{\theta}$ are obtained by maximizing $\tilde{L}_{S,m}(\underline{\theta};\,\underline{y},\underline{\hat{\alpha}})$ with respect to $\underline{\theta}$, using an iterative numerical optimizer. The convergence of such an optimizer requires the use of CRNs in order to ensure that $\tilde{L}_{S,m}(\underline{\theta};\,\underline{y},\underline{\hat{\alpha}})$ is a smooth function in $\underline{\theta}$.

13.6.3 Empirical Application

We applied the EIS algorithm described above to the ML estimation of a panel logit model for the union/non-union decision of young men. This application is based upon the framework and dataset used in the study of Vella and Verbeek (1998).

In particular, the reduced form model for the choice of individual i in period t of a union or a non-union employment is assumed to be

$$y^*_{ti} = \underline{\beta}'\underline{z}_{ti} + \gamma y_{t-1,i} + \tau_i + \lambda_t + \varepsilon_{ti}\,, \qquad i:1\rightarrow N, \quad t:2\rightarrow T, \tag{13.76}$$

where the latent variable y^*_{ti} represents the individual benefits of a union membership. The observed union status is given by the dummy $y_{ti} = I_{\{y^*_{ti}\geq 0\}}$. As proposed by Vella and Verbeek (1998) the union status is explained by individuals' characteristics and a set of industry dummies summarized in the vector $\underline{z}_{ti}$ and by the lagged union status $y_{t-1,i}$. The lagged union status is included to capture individuals' propensity to remain in the initially chosen status. For simplicity, we assume that the initial state $y_{1,i}$ is a fixed non–stochastic constant for individual i. (A more sophisticated alternative to handle this initial condition problem proposed by Heckman (1981) and followed by Vella and Verbeek (1998) is to approximate the marginal probability of $y_{1,i}$ using all information on the exogenous variables in period $t = 1$.) For the time random effect λ_t we use a first–order autoregressive process.

In contrast to this panel logit specification with random individual and random dynamic time effects, Vella and Verbeek (1998) employ a corresponding probit model with random individual and fixed time effects. Here a logit specification is used just for computational convenience, but a substitution of the logistic distribution for ε_{ti} by a normal distribution or any other suitable distribution requires only minor modifications in the EIS algorithm. The use of random individual and fixed time effects enables Vella and Verbeek (1998) to rely on standard integration procedures for one–dimensional integrals to evaluate the likelihood function, but their procedure cannot be applied to evaluate the likelihood for random individual and random dynamic time effects jointly. By applying EIS we can do the latter and compare both approaches.

The data used to estimate the model are taken from the National Longitudinal Survey (NLS Youth Sample) and contain observations on 545 males for the years 1980–1987 (for a detailed description of the data, see Vella and Verbeek (1998)). The ML EIS estimates of the model based upon a simulation sample size $S = 300$ and three iterations of the efficient sampling algorithm are given in Table 13.1. Each likelihood evaluation requires approximately 2 seconds on a Pentium 4, 3.06 GHz personal computer for a code written in GAUSS. A full ML estimation requires approximately 122 BFGS iterations and takes of the order of 230 min. The parameter estimates are numerically accurate, as indicated by the MC (numerical) standard errors, which were computed from 20 ML estimations conducted under different sets of CRNs.

The parameter estimates for the impact of the individuals' characteristics and the industry dummies on the probability of union membership are consistent with those reported by Vella and Verbeek (1998). In particular, except for the variable $\log(1 + \text{experience})$, whose estimated impact is not significant, the signs of the parameter estimates are all the same in both estimates. To make the values of the parameter estimates from our logit specification comparable with those from Vella and Verbeek's probit specification, one can divide our parameter estimates by $(\pi^2/3 + \hat{\sigma}_\tau^2 + \hat{\sigma}_\lambda^2)^{1/2}$. The result (not presented here) shows that the values are for all parameters very close together. Furthermore, only for the variables *health disability*, *Lives in North East*, and *Entertainment* our model estimate leads to different conclusions with respect to statistical significance compared to Vella and Verbeek's estimate.

The estimate of the variance parameter of the random individual effects σ_τ is significantly greater than zero and its estimate of 1.77 indicates that 48.6% of total variation is explained by cross individual variation which is lower than the 57% reported by Vella and Verbeek (1998). Furthermore, the estimate of the variance parameter σ_λ is 2.19 standard errors larger than zero and implies that 0.5% of the total variance is explained by random time effects. Finally, observe that the autoregressive coefficient is not significantly different from zero. Together with the fact that lagged union status has a highly significantly positive impact on the probability of union membership this indicates that the dynamics in the union/non-union decision is dominated by an idiosyncratic component rather than by an aggregate common component.

Table 13.1 ML efficient sampling estimates of the panel logit model for union membership

Variable	Estimate	asy. std. error	MC std. error
Constant	−1.1912	1.2607	0.1373
Lagged union status	1.9685	0.1483	0.0565
log(1+experience)	−0.2523	0.3075	0.0439
Years of schooling	−0.0385	0.0595	0.0063
Married	0.3408	0.1430	0.0080
Black	1.2835	0.2722	0.0354
Hispanic	0.6267	0.2209	0.0280
Lives in rural area	0.0071	0.1901	0.0148
Has health disability	−0.6448	0.1934	0.0123
Lives in North East	0.4194	0.2348	0.0229
Lives in south	−0.0593	0.2164	0.0234
Lives in Northern Central	0.3996	0.2265	0.0159
Industry Dummies			
Agricultural	−1.4372	0.4450	0.0294
Mining	−0.6509	0.4995	0.0355
Construction	−1.1622	0.3498	0.0229
Manufacturing	−0.5519	0.3044	0.0141
Transportation	−0.2467	0.3547	0.0195
Trade	−1.4442	0.3127	0.0169
Finance	−3.0984	0.5065	0.0902
Business & repair service	−2.0654	0.3880	0.0207
Personal service	−2.0703	0.3936	0.0219
Entertainment	−1.5235	0.4227	0.0324
Professional & related services	−0.4990	0.3299	0.0149
Random Effects Parameters			
σ_τ	1.7735	0.1367	0.0777
σ_λ	0.1774	0.0809	0.0154
ρ_1	−0.1124	0.5171	0.1690
Log-likelihood value	−1303.71		2.1713

Note: Asymptotic standard errors are obtained from a numerical approximation to the Hessian. The ML efficient sampling estimates are based on a MC sample sice of $S = 300$ and three iterations of the construction of the efficient sampler.

13.7 Conclusion

Simulation based inference procedures have become a key component of the micro– and macroeconometrician's toolbox. This chapter was never meant to provide a systematic survey of the recent literature. As mentioned in the course of the discussion, excellent surveys are available elsewhere (and often require more than a single chapter!). See, in particular, Gourieroux and Monfort (1993, 1994). We have attempted to selectively discuss issues which, based upon our own experience, constitute the cornerstones of an efficient usage of MC simulation techniques, with emphasis on efficient integration of random effects in panel models.

Much of our discussion applies to classical and Bayesian procedures as well, largely because we insist on interpreting MC integration as a *numerical* technique for constructing approximations to expression which depend upon (high-dimensional) integrals for which no analytical expressions are available.

We find ourselves at odds with the recent (classical) literature on simulation based estimation on two key counts.

Firstly, "natural" MC simulation is increasingly outdated and often utterly impractical in applications requiring moderate to high–dimensional interdependent integration (especially for "marginalized" likelihood evaluation). There exists an increasing range of operational "acceleration" procedures, most prominently Efficient Importance Sampling, which can produce considerable efficiency gains which generally far outweighs moderate increases in the cost of computations for any given number of draws. Yet acceleration procedures are largely ignored by the classical literature.

Secondly, one ought to draw a clear distinction between the *statistical* properties of an estimator and the *numerical* accuracy of its simulated counterpart. Unsurprisingly, the implementation of that key distinction greatly simplifies the conceptual and practical analysis of simulation based estimators.

Acknowledgments Financial support for this work has been provided by the National Science Foundation (SES–9223365). We are especially indebted to Jean–Pierre Florens, John Geweke, David Hendry, Jay Kadane and Wei Zhang for numerous helpful discussions relative to the use of simulation techniques in econometrics.

13.8 Appendix: Implementation of EIS for the Logit Panel Model

Implementation of the efficient sampling procedure for the likelihood evaluation for the panel logit model (13.59)–(13.60) starts with the selection of the class of density kernels k_i for the auxiliary samplers m_i capable of approximating ϕ_i as defined in (13.68). Since the natural sampler for τ_i is a constituent component of ϕ_i, a natural choice for m_i is, as mentioned above, a parametric extension of the natural sampler. In our case, this leads to a Gaussian density kernel k_i for $\tau_i \mid \underline{\lambda}$. In particular, the following parametrization is used:

$$k_i(\tau_i;\, \underline{\lambda}, \underline{\alpha}_i) = \exp\left\{ -\frac{1}{2}\left(\underline{b}_i' \underline{v}_i + \underline{v}_i' C_i \underline{v}_i\right) - \frac{\tau_i^2}{2\sigma_\tau^2} \right\}, \tag{13.77}$$

where

$$\underline{b}_i = (b_{1i}, \ldots, b_{Ti})', \quad C_i = \operatorname{diag}(\underline{c}_i), \quad \underline{c}_i = (c_{1i}, \ldots, c_{Ti})' \tag{13.78}$$

$$\underline{v}_i = \underline{\lambda} + \tau_i \underline{\iota} + Z_i \underline{\beta}, \quad \text{with} \quad \underline{\iota} = (1, \ldots, 1)', \quad Z_i = \left(\underline{z}_{1i}, \ldots, \underline{z}_{Ti}\right)', \tag{13.79}$$

and the auxiliary parameters are $\underline{\alpha}_i = (\underline{b}_i', \underline{c}_i')'$. Note that under this parametrization of k_i the factor $\exp\{-\tau_i^2/(2\sigma_\tau^2)\}$ cancels out in the LS problems (13.74). In order

to derive the conditional mean and variance of the Gaussian sampling density m_i, it is useful to rewrite k_i as follows:

$$k_i(\tau_i;\ \underline{\lambda}, \underline{\alpha}_i) = \exp\left\{-\frac{1}{2}\left[\left(\frac{1}{\sigma_\tau^2} + \underline{\iota}'C_i\underline{\iota}\right)\tau_i^2 + \left(\underline{b}_i'\underline{\iota} + 2\underline{\iota}'C_i\underline{\ell}_i\right)\tau_i + \underline{b}_i'\underline{\ell}_i + \underline{\ell}_i'C_i\underline{\ell}_i\right]\right\}, \tag{13.80}$$

where $\underline{\ell}_i = \underline{\lambda} + Z_i\underline{\beta}$. Accordingly, the conditional mean and variance of $\tau_i|\underline{\lambda}$ on m_i are obtained as

$$\mu_i = -\sigma_i^2\left(\frac{1}{2}\underline{b}_i'\underline{\iota} + \underline{\iota}'C_i\underline{\ell}_i\right) \quad \text{and} \quad \sigma_i^2 = \frac{\sigma_\tau^2}{1 + \underline{\iota}'C_i\underline{\iota}\sigma_\tau^2}\,. \tag{13.81}$$

Integrating k_i with respect to τ_i leads to the following form of the integrating constant:

$$\chi_i(\underline{\lambda}, \underline{\alpha}_i) \propto \exp\left\{-\frac{1}{2}\left[\underline{b}_i'\underline{\ell}_i + \underline{\ell}_i'C_i\underline{\ell}_i - \frac{\mu_i^2}{\sigma_i^2}\right]\right\}. \tag{13.82}$$

which itself is a Gaussian density kernel for $\underline{\lambda}$. Based on these functional forms, the computation of an efficient MC estimate of the likelihood for the panel logit model requires the following steps:

Step (1): Use the natural sampling density p to draw S independent realizations of the latent processes $(\underline{\tilde{\tau}}_r, \underline{\tilde{\lambda}}_r)$.

Step (2): Use these random draws to solve the sequence of N weighted (unweighted for the first iteration of importance sampling construction) LS problems defined in (13.74). The ith weighted LS problem is characterized by the following linear auxiliary regression:

$$\begin{aligned}\sum_{t=1}^{T}\left[y_{ti}\tilde{v}_{tir} - \ln\left(1 + \exp\{\tilde{v}_{tir}\}\right)\right] = \text{constant} + (-b_{1i}/2)\,\tilde{v}_{1ir} + \cdots + (-b_{Ti}/2)\,\tilde{v}_{Tir} \\ + (-c_{1i}/2)\,\tilde{v}_{1ir}^2 + \cdots + (-c_{Ti}/2)\,\tilde{v}_{Tir}^2 + \xi_{ir}\,,\end{aligned} \tag{13.83}$$

with weights:

$$g_i(\underline{\tilde{\tau}}_r, \underline{\tilde{\lambda}}_r;\ \underline{\theta}, \underline{y}) = \prod_{t=1}^{T}\left[\frac{1}{1 + \exp\{\tilde{v}_{1ir}\}}\right]^{1-y_{ti}}\left[\frac{\exp\{\tilde{v}_{1ir}\}}{1 + \exp\{\tilde{v}_{1ir}\}}\right]^{y_{ti}}, \tag{13.84}$$

where ξ_{ir} denotes the regression error term and $\{\tilde{v}_{tir};\ r : 1 \to S\}$ are the simulated draws of v_{ti}.

Step (3): The function to be approximated by the Gaussian sampler m_0 is given by:

$$\phi_0(\underline{\lambda};\ \underline{\theta})\prod_{i=1}^{N}\chi_i(\underline{\lambda}_i, \underline{\alpha}_i) \propto \exp\left\{-\frac{1}{2}\left(\underline{\lambda}'\Sigma_\lambda^{-1}\underline{\lambda} + \sum_{i=1}^{N}\left[\underline{b}_i'\underline{\ell}_i + \underline{\ell}_i'C_i\underline{\ell}_i - \frac{\mu_i^2}{\sigma_i^2}\right]\right)\right\}, \tag{13.85}$$

which has as mentioned above the form of a Gaussian kernel for $\underline{\lambda}$. Accordingly, the mean and variance-covariance matrix of $\underline{\lambda}$ on m_0 are obtained as

$$\underline{\mu}_0 = \Sigma_0 \sum_{i=1}^{N} \left[\sigma_i^2 \underline{c}_i \left(\underline{c}_i' Z_i \underline{\beta} + \frac{1}{2} \underline{\iota}' \underline{b}_i \right) - \frac{1}{2} \underline{b}_i - C_i Z_i \underline{\beta} \right] \tag{13.86}$$

$$\Sigma_0 = \left[\Sigma_\lambda^{-1} + \sum_{i=1}^{N} \left(C_i - \sigma_i^2 \underline{c}_i \underline{c}_i' \right) \right]^{-1} . \tag{13.87}$$

Use this sampling density m_0 to draw S trajectories $\{\tilde{\underline{\lambda}}_r(\hat{\underline{\alpha}}_0);\ r : 1 \to S\}$. Conditional on these trajectories, draw from the conditional densities $\{m_i\}$ characterized by the moments (13.81) the vectors $\{\tilde{\underline{\tau}}_r(\hat{\underline{\alpha}}_1, \ldots, \hat{\underline{\alpha}}_N);\ r : 1 \to S\}$. Based on the draws $\{(\tilde{\underline{\tau}}_r(\hat{\underline{\alpha}}_1, \ldots, \hat{\underline{\alpha}}_N);\ \tilde{\underline{\lambda}}_r(\hat{\underline{\alpha}}_0))\}$ the efficient sampling estimate of the likelihood is calculated according to (13.71).

References

Bauwens, L. and N. Hautsch (2003), Stochastic Conditional Intensity Process, *Mimeo*, Universite Catholique de Louvain.

Bertschek, I. (1995), Product and Process Innovation as a Response to Increasing Imports and Foreign Direct Investment, *Journal of Industrial Economics*, 43, pp. 341–357.

Bertschek, I. and M. Lechner (1998), Convenient Estimators for the Panel Probit Model, *Journal of Econometrics*, 87, pp. 329–372.

Brown, B.W., A. Monfort and H.K. Van Dijk (1993), Econometric Inference using Simulation Techniques, *Journal of Applied Econometrics*, 8, pp. s1–s3 (special issue).

Börsch–Supan, A. and V. Hajivassiliou (1993), Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models, *Journal of Econometrics*, 58, pp. 347–368.

Börsch–Supan, A., V. Hajivassiliou, L. Kotlikoff and J. Morris (1990), Health, Children, and Elderly Living Arrangements: A Multiperiod-Multinomial Probit Model with Unobserved Heterogeneity and Autocorrelated Errors, *National Bureau of Economic Research*, Working Paper No. 3343.

Danielsson, J. and J.F. Richard (1993), Accelerated Gaussian Importance Sampler with Application to Dynamic Latent Variable Models, *Journal of Applied Econometrics*, 8, pp. 153–173.

Davidson, R. and J.G. McKinnon (1992), Regression–based Methods for Using Control Variates in Monte Carlo Experiments, *Journal of Econometrics*, 54, pp. 203–222.

Davidson, R. and J.G. McKinnon (1993), *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

Devroye, L. (1986), *Non–Uniform Random Variate Generation*, Springer–Verlag, New York.

Fishman, G.S. (1996), *Monte Carlo Concepts, Algorithms, and Applications*, Springer, New York.

Fishman, G.S. and L.R. Moore, III (1982), A Statistical Evaluation of Multiplicative Random Number Generators with Modulus $2^{31} - 1$, *Journal of the American Statistical Association*, 77, pp. 129–136.

Fishman, G.S. and L.R. Moore, III (1986), An Exhaustive Analysis of Multiplicative Congruential Random Number Generators with Modulus $2^{31} - 1$, *SIAM Journal on Scientific and Statistical Computing*, 7, pp. 24–45.

Gallant, A.R. (1987), Nonlinear Statistical Models, John Wiley & Sons, New York.

Geweke, J. (1988), Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference, *Journal of Econometrics*, 38, pp. 73–89.

Geweke, J. (1989), Bayesian Inference in Econometric Models Using Monte Carlo Integration, *Econometrica*, 57, pp. 1317–1340.

Geweke, J. (1991), Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints, *Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, pp. 571–578.

Geweke, J. (1994), Monte Carlo Simulation and Numerical Integration, *Federal Reserve Bank of Minneapolis*, Working Paper No. 526.

Geweke, J., M. Keane and D. Runkle (1994), Alternative Computational Approaches to Inference in the Multinomial Probit Model, *Review of Economics and Statistics*, 76, pp. 609–632.

Geweke, J., M. Keane and D. Runkle (1997), Statistical Inference in the Multinomial Multiperiod Probit Model, *Journal of Econometrics*, 81, pp. 125–166.

Greene, W. (2004), Convenient Estimators for the Panel Probit Model: Further Results, *Empirical Economics*, 29, pp. 21–47.

Gourieroux, C. and A. Monfort (1993), Simulation Based Inference: A Survey with Special Reference to Panel Data Models, *The Journal of Econometrics*, 59, pp. 5–33.

Gourieroux, C. and A. Monfort (1994), Simulation Based Econometric Methods, *CORE Lecture Series*, CORE Foundation, Louvain–la–Neuve.

Hajivassiliou, V. (1990), *Smooth Simulation Estimation of Panel Data LDV Models*, Mimeo, Yale University.

Heckman, J.J. (1981), The Incidental Parameters Problem and the Problem of Initial Conditions in Estimation a Discrete Time–Discrete Data Stochastic Process, in *Structural Analysis of Discrete Data with Econometric Applications (chapter 4) Manski, C.F. and D. McFadden,* MIT Press, Cambridge.

Hendry, D.F. (1984), Monte Carlo Experimentation in Econometrics, in *The Handbook of Econometrics (chapter 16) Griliches, Z. and M. Intriligator,* North-Holland, Amsterdam.

Hendry, D.F. and J.F. Richard (1992), Likelihood Evaluation for Dynamic Latent Variable Models, in *Computational Economcis and Econometrics (chapter 1) Amman, H.M, Belsley, D.A. and C.F. Pau*, Kluwer Academic Publishers, Dordrecht.

Jung, R.C. and R. Liesenfeld (2001), Estimating Time Series Models for Count Data Using Efficient Importance Sampling, *Allgemeines Statistisches Archiv*, 85, pp. 387–407.

Keane, M. (1994), A Computationally Practical Simulation Estimator for Panel Data, *Econometrica*, 62, pp. 95–116.

Kloek, T. and H.K. van Dijk (1978), Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo, *Econometrica*, 46, pp. 1–19.

Lerman, S. and C. Manski (1981), On the Use of Simulated Frequencies to Approximate Choice Probability, in *Structural Analysis of Discrete Data with Econometric Applications (chapter 7) C. Manski and D. McFadden*, MIT Press, Cambridge.

Liesenfeld, R. and J.F. Richard (2003a), Monte Carlo Methods and Bayesian Computation: Importance Sampling, in *The International Encyclopedia of the Social and Behavioral Sciences (pp. 10000–10004) Smelser, N.J. and P.B. Baltes*, Elsevier Science, Oxford.

Liesenfeld, R. and J.F. Richard (2003b), Estimation of Dynamic Bivariate Mixture Models: Comments on Watanabe (2000), *The Journal of Business and Economic Statistics*, 21, pp. 570–576.

Liesenfeld, R. and J.F. Richard (2003c), Univariate and Multivariate Volatility Models: Estimation and Diagnostics, *The Journal of Empirical Finance*, 10, pp. 505–531.

McFadden, D. (1989), A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration, *Econometrica*, 57, pp. 995–1026.

Pakes, A. and D. Pollard (1989), Simulation and the Asymptotics of Optimization Estimators, *Econometrica*, 57, pp. 1027–1058.

Press, W.h., B.P. Flannery, S.A. Teukolsky and W.T. Vetterling (1986), *Numerical Recipes (The Art of Scientific Computing)*, Cambridge University Press, Cambridge.

Richard, J.F. and W. Zhang (2007), Efficient High-Dimensional Importance Sampling, Journal of Econometrics, 141, pp. 1385–1411.

Robert, C.P. and G. Casella (2004), *Monte Carlo Statistical Methods*, Springer, New York.
Stern, S. (1992), A Method for Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models, *Econometrica*, 60, pp. 943–952.
Stern, S. (1997), Simulation-Based Estimation, *Journal of Economic Literature*, 35, pp. 2006–2039.
Stat/Library User's Manual (1991) IMSL, IMSL Inc., Houston.
Vella, F. and M. Verbeek (1998), Whose Wages do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men, *Journal of Applied Econometrics*, 13, pp. 163–283.

Chapter 14
Semi-parametric and Non-parametric Methods in Panel Data Models

Chunrong Ai and Qi Li

14.1 Introduction

Common approach for modeling panel data $\{(y_{it}, x_{it}), i = 1, 2, \ldots, N; t = 1, 2, \ldots, T\}$ involves postulating that the data generating process depends on a time invariant individual specific effect α_i and some model parameters, and satisfies some statistical restrictions. The individual effect α_i is not observed and is assumed to be randomly distributed across individuals. The model parameter may include finite dimensional as well as infinite dimensional parameter. Estimation of the parameter of interest depends on the statistical restrictions imposed on the data generating process and on the relative values of N and T. Since many different restrictions can be imposed on the data generating process and different restrictions often lead to different estimation methods, the econometric literature on estimation of panel data models is large and is scattered around various academic outlets. The aim of this handbook is to provide a thorough survey of the vast literature on estimation of panel data models. Within the general objective of the handbook, this chapter attempts to survey recent development on estimation methods of a particular class of models: semiparametric and nonparametric panel data models. For discussions on parametric panel data models, see Baltagi (2005) and Hsiao (2003).

A model is semiparametric or nonparametric if the model parameter includes the infinite dimensional parameter. By this definition, the classical linear panel data model

$$y_{it} = x_{it}'\theta_o + \alpha_i + u_{it}\ , \qquad i = 1, 2, \ldots, N;\ t = 1, 2, \ldots, T\ , \tag{14.1}$$

Chunrong Ai
Department of Economics, University of Florida, Warrington College of Business Administration, 224 MAT, P.O. Box 117140, Gainesville, FL 32611-7140, USA, e-mail: chunrong.ai@cba.ufl.edu

Qi Li
Department of Economics, Texas A&M University, College Station, TX 77843-4228, USA, e-mail: qi@econmail.tamu.edu

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

is a semiparametric model if the joint distribution of $(x_{it}, \alpha_i, u_{it})$ is not parameterized. In fact, by this definition, almost all panel data models that have been studied in the literature, including those discussed in other chapters of this handbook, can be classified as the semiparametric models. Thus, surveying the literature on this general class of semiparametric and nonparametric models will duplicate the discussions in the other chapters of the handbook. Our focus here is instead on a smaller class of semiparametric and nonparametric panel data models that are not covered by the other chapters. Specifically, we will survey the literature on: (1) partly linear or fully nonparametric panel data models and (2) panel data discrete choice and censored regression models with unknown error distribution. We note that Arellano and Honore (2001) reviewed the panel data discrete choice and censored regression models. Although we will survey those same models, we will include more recent results as well as various extensions to those models.

Throughout the chapter, we will use $i = 1, 2, \ldots, N$ to denote an individual and $t = 1, 2, \ldots, T$ to denote time. T is small relative to N. We will restrict our discussion to the balanced panel (i.e., T does not vary with i) for the purpose of simplifying exposition. The methods we review here are easily applicable to the unbalanced panel. We will use $1\{A\}$ to denote the indicator function that takes the value 1 if event A occurs and the value 0 otherwise. $sgn(\cdot)$ is the sign function that takes the value 1 if $\cdot > 0$, the value 0 if $\cdot$ is zero, and the value -1 if $\cdot$ is negative. For each i, denote $x_i = (x_{i1}, x_{i2}, \ldots, x_{iT})$, $y_i = (y_{i1}, \ldots, y_{iT})$, and $u_i = (u_{i1}, \ldots, u_{iT})$. Finally, variables with superscript $*$ denote the latent variables that are not always observed directly.

14.2 Linear Panel Data Model

14.2.1 Additive Effect

We begin with the classical linear panel data model given by (14.1), where the individual effect enters the model additively and x_{it} does not include the time-invariant regressors, w_i. The linear and the additive structure of the model allows us to remove the individual effect through mean-differencing:

$$y_{it} - \frac{1}{T}\sum_{s=1}^{T} y_{is} = (x_{it} - \frac{1}{T}\sum_{s=1}^{T} x_{is})'\theta_o + u_{it} - \frac{1}{T}\sum_{s=1}^{T} u_{is} .$$

A common assumption for this model is that the explanatory variables satisfy the following strict exogeneity condition:

Assumption 1 (Strictly exogenous regressors). $E(u_i|x_i) = 0$.
Under this condition, the coefficient on the time-variant regressors, θ_o, can be estimated by the fixed effects (within) estimator given by

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N}\sum_{t=1}^{T} \left(y_{it} - \frac{1}{T}\sum_{s=1}^{T} y_{is} - (x_{it} - \frac{1}{T}\sum_{s=1}^{T} x_{is})'\theta \right)^2 .$$

The fixed effects estimator $\widehat{\theta}$ is consistent and asymptotically normally distributed under some standard regularity conditions. The fixed effects estimator $\widehat{\theta}$ is inconsistent, however, if the model is misspecified. The model is misspecified, for example, if the deterministic part $x_{it}'\theta_o$ is incorrectly specified or if Assumption 1 is not satisfied. To avoid potential specifications, one can use more flexible specifications. For instance, Robinson's (1988) partly linear function $x_{1it}'\theta_o + h_o(x_{2it})$ is more flexible than the linear function $x_{it}'\theta_o$; the additive nonparametric function $h_{1o}(x_{1it}) + h_{2o}(x_{2it})$ is more flexible than the partly linear function; and the nonparametric function $h_o(x_{it})$ is the most flexible specification. All of these specifications can be nested in the following general index model:

$$y_{it} = v_0(x_{it}, \theta_o) + \sum_{j=1}^{m} h_{j0}(v_j(x_{it}, \theta_o)) + \alpha_i + u_{it} \tag{14.2}$$

where $v_j(\cdot)$ for $j = 0, 1, \ldots, m$ are known functions and $h_{j0}(\cdot) = 0$ for $j = 1, 2, \ldots, m$ are unknown functions. The parameter of interest now includes the finite dimensional parameter θ_o and the infinite dimensional parameter

$$h_o(\cdot) = (h_{1o}(\cdot), \cdots, h_{mo}(\cdot)) .$$

Identification of model (14.2) may require some restrictions on $(\theta_o, h_o(\cdot))$. Notice that it is impossible to distinguish $(h_{so}(\cdot), \alpha_i)$ from $(h_{so}(\cdot) - \mu, \alpha_i + \mu)$ for any constant μ and for any s. Therefore, identification of the model parameter requires some location restrictions such as $h_{jo}(0) = 0$ for all j. Identification of the model parameter may also require scaling restrictions when the function $v_j(x, \theta)$ for some $j > 0$ is homogenous of degree one in θ. To illustrate, consider the simpler model

$$y_{it} = h_o(x_{it}'\theta_o) + \alpha_i + u_{it} .$$

Evidently we cannot distinguish $(\theta_o, h_o(\cdot))$ from $(\widetilde{\theta}_o, \widetilde{h}_o(\cdot))$, with $\widetilde{h}_o(\cdot) = h_o(\cdot/\mu)$ and $\widetilde{\theta}_o = \theta_o\mu$, for any nonzero constant μ. On the other hand, a scaling restriction such as $\theta_o'\theta_o = 1$ or $\theta_o = (1, \theta_{2o}, \ldots, \theta_{ko})'$ if the first coefficient in θ_o is nonzero distinguishes $(\theta_o, h_o(\cdot))$ from $(\widetilde{\theta}_o, \widetilde{h}_o(\cdot))$ and hence should be imposed for identification purpose. Finally, identification of the model parameter may require exclusion restrictions when $v_j(x, \theta)$ and $v_s(x, \theta)$ for some $s \neq j$ are homogenous of degree one in the regressors. To see this, consider

$$y_{it} = h_{1o}(x_{1it}) + h_{2o}(x_{2it}) + \alpha_i + u_{it} .$$

Clearly (h_{1o}, h_{2o}) is not distinguishable from $(h_{1o} + g(x_{3it}), h_{2o} - g(x_{3it}))$ for any function g if x_{1it} and x_{2it} contain the common regressor x_{3it}. In this case, an exclusive restriction such as that x_{1it} and x_{2it} are mutually exclusive distinguishes (h_{1o}, h_{2o}) from $(h_{1o} + g(x_{3it}), h_{2o} - g(x_{3it}))$ and hence must be imposed.

Suppose that the parameter of interest is identified up to some location restrictions. Under Assumption 1, a natural approach for estimating (θ_o, h_o) would be to apply the fixed effects estimation. The problem with this approach, however, is

that h_o is infinite dimensional and cannot be estimated from the finite data points. One can use various nonparametric estimation techniques to estimate the unknown function h_o such as kernel, nearest neighbor and series method. Model (14.2) has an additive structure. It is known that the nonparametric series approach is the most convenient to estimate restricted nonparametric regression models such as additive, multiplicative and monotonic restrictions. Therefore, in this chapter we will mainly focus on the series estimation method. For using kernel marginal integration method to estimate additive models, see Linton and Nielsen (1995), Newey (1994), and Tjostheim and Auestad (1994).

One can replace the unknown functions with finite dimensional series approximations and then apply the fixed effects estimation to the model as if the finite dimensional approximations are the correct specifications. Specifically, for each j, let $p_j(\cdot) = (p_{j1}(\cdot), q_{j2}(\cdot), \ldots, p_{jk_j}(\cdot))'$ denote known basis functions that approximate $h_{jo}(\cdot)$ in the sense that, for some constant vector π_j of dimension $k_j \times 1$ and a scalar constant $\tau_j > 0$,

$$h_{jo}(\cdot) = p_j(\cdot)'\pi_j + O(k_j^{-\tau_j}) = h_{k_j j}(\cdot) + O(k_j^{-\tau_j}) .$$

The simplest series base function is the power series, $\{1, x, x^2, \ldots\}$. However, estimation based on power series can be sensitive to outliers and, for this reason, the power series is not typically used for nonparametric series estimation. Instead, the piecewise local polynomial spline is the most commonly used base function in nonparametric series estimation. An rth order univariate B-spline base function is given by (see Chui (1992, Chap. 4))

$$B_r(x|t_0, \ldots, t_r) = \frac{1}{(r-1)!} \sum_{j=0}^{r} (-1)^j \binom{r}{j} [\max(0, x - t_j)]^{r-1} , \tag{14.3}$$

where $t_0, \ldots, t_r$ are the evenly spaced design knots on the support of X, and a is the distance between knots. When $r = 2$ (14.3) gives a piecewise linear spline, and when $r = 4$, it gives piece-wise cubic splines (i.e., third order polynomials).

In finite sample applications, the approximating function $h_{k_j j}(\cdot)$ is obviously finite dimensional, depending on k_j unknown coefficients π_j. If each $h_{jo}(\cdot)$ is replaced by $h_{k_j j}(\cdot)$, the total number of unknown coefficients in the approximating functions is $k = k_1 + \ldots + k_m$. Let d_θ denote the dimension of θ and denote $\pi = (\pi_1', \ldots, \pi_m')'$. Clearly, in order to estimate (θ_o, π), the total number of coefficients to be estimated must be smaller than the sample size: $k + d_\theta < NT$. Moreover, the approximating functions must satisfy the location restriction $h_{k_j j}(0) = p_j(0)'\pi_j = 0$. Notice that it is always possible to choose $p_j(\cdot)$ so that $p_j(0) = 0$. Without loss of generality, we will assume that the basis functions $p_j(\cdot)$ satisfy $p_j(0) = 0$. Hence, there is no need to impose restriction on π_j. Now treat each approximation $p_j(\cdot)'\pi_j$ as if it is the correct specification of $h_{jo}(\cdot)$. The fixed effects estimator (θ_o, π) is given by

$$\begin{pmatrix} \widehat{\theta} \\ \widehat{\pi} \end{pmatrix} = \arg \min_{\theta, \pi_j, 1 \le j \le m} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\begin{array}{c} y_{it} - v_0(x_{it}, \theta) - \sum_{j=1}^{m} p_j(v_j(x_{it}, \theta))'\pi_j - \\ \frac{1}{T} \sum_{s=1}^{T} (y_{is} - v_0(x_{is}, \theta) - \sum_{j=1}^{m} p_j(v_j(x_{is}, \theta))'\pi_j) \end{array} \right]^2 . \tag{14.4}$$

The unknown functions $h_{jo}(\cdot)$ for $j = 1,\ldots,m$ are estimated by $\widehat{h}_j(\cdot) = p_j(\cdot)'\widehat{\pi}_j$. Obviously, in order for the fixed effects estimator to be consistent, the approximation errors must shrink to zero as sample size goes to infinity. This can happen if we require each k_j goes to infinity as $N \to \infty$ but at a slower rate so that $\frac{k_j}{N} \to 0$. Shen (1997), Newey (1997) and Chen and Shen (1998) show that both $\widehat{\theta}$ and $\widehat{h}_j(\cdot)$, $j = 1,2,\ldots,m$ are consistent. They also show that, if $k \to \infty$ at certain rate, the estimator $\widehat{\theta}$ is root-N consistent and asymptotically normally distributed, and any smooth functional of $\widehat{h}_j(\cdot)$ for $j = 1,2,\ldots,m$, is asymptotically normally distributed. Moreover, Shen (1997) shows that the usual covariance matrix of the fixed effects estimator $\widehat{\theta}$ is a consistent estimator of the asymptotic variance-covariance of $\widehat{\theta}$. Denote $\delta = (\theta', \pi')'$ with $\widehat{\delta} = (\widehat{\theta}', \widehat{\pi}')'$ and denote

$$l_1(y_i, x_i, \delta) = \sum_{t=1}^{T} \left[\begin{array}{c} y_{it} - v_0(x_{it}, \theta) - \sum_{j=1}^{m} p_j(v_j(x_{it}, \theta))'\pi_j - \\ \frac{1}{T}\sum_{s=1}^{T}(y_{is} - v_0(x_{is}, \theta) - \sum_{j=1}^{m} p_j(v_j(x_{is}, \theta))'\pi_j) \end{array} \right]^2 . \quad (14.5)$$

Denote

$$\begin{aligned} \widehat{V} &= \left(\sum_{i=1}^{N} \frac{\partial^2 l_1(y_i, x_i, \widehat{\delta})}{\partial \delta \partial \delta'} \right)^{-1} \times \left(\sum_{i=1}^{N} \frac{\partial l_1(y_i, x_i, \widehat{\delta})}{\partial \delta} \frac{\partial l_1(y_i, x_i, \widehat{\delta})}{\partial \delta'} \right) \\ &\quad \times \left(\sum_{i=1}^{N} \frac{\partial^2 l_1(y_i, x_i, \widehat{\delta})}{\partial \delta \partial \delta'} \right)^{-1} \\ &= \left(\begin{array}{cc} \widehat{V}_\theta & \widehat{V}_{\theta\pi} \\ \widehat{V}_{\pi\theta} & \widehat{V}_\pi \end{array} \right) . \end{aligned}$$

Then, $\widehat{V}$ is the covariance matrix of the fixed effects estimator $\widehat{\delta}$ if $h_{k_j j}(\cdot)$ is the correct specification of $h_{jo}(\cdot)$ for fixed value of k. $\widehat{V}$ is not the covariance matrix of $\widehat{\delta}$ when $k \to \infty$ because $P;(\cdot)'\widehat{\pi}$; is not a root-N consistent estimator for $h_{jo}(\cdot)$. But the upper-left block $\widehat{V}_\theta$ is a consistent estimator for the covariance matrix of $\widehat{\theta}$.

The fixed effects estimator $\widehat{\delta}$ can be used for testing model specifications. For example, applying the idea of Hong and White (1995), one can develop a consistent test for the more restrictive model (14.1) against the general index model (14.2). Specifically, let $\widetilde{\theta}$ denote the fixed effects estimator for model (14.1). The test statistic is constructed by comparing the fitted values under both models:

$$\begin{aligned} \widehat{\lambda} = \sum_{i=1}^{N}\sum_{t=1}^{T} &\left[y_{it} - x_{it}'\widetilde{\theta} - \frac{1}{T}\sum_{s=1}^{T}(y_{is} - x_{is}'\widetilde{\theta}) \right] * \\ &\left[\begin{array}{c} v_0(x_{it}, \widehat{\theta}) + \sum_{j=1}^{m} p_j(v_j(x_{it}, \widehat{\theta}))'\widehat{\pi}_j - x_{it}'\widetilde{\theta} \\ -\frac{1}{T}\sum_{s=1}^{T}(v_0(x_{is}, \widehat{\theta}) + \sum_{j=1}^{m} p_j(v_j(x_{is}, \widehat{\theta}))'\widehat{\pi}_j - x_{it}'\widetilde{\theta}) \end{array} \right] . \end{aligned}$$

The asymptotic distribution of $\widehat{\lambda}$ can be established with techniques similar to those developed in Hong and White (1995), Fan and Li (1996), and Li and Wang (1998). Applying the same idea, one can also develop a consistent test for the partly specified model (14.2) against the fully nonparametric panel data model:

$$y_{it} = g(x_{it}) + \alpha_i + u_{it} \ , \tag{14.6}$$

where $g(\cdot)$ is an unknown function satisfying $g(0) = 0$. Suppose that $g(\cdot)$ is approximated by the basis functions $q(x_{it}) = (q_1(x_{it}), \ldots, q_k(x_{it}))'$ in the sense that there exist $k \times 1$ vector β and a constant $\tau > 0$ such that

$$g(x_{it}) = q(x_{it})'\beta + O(k^{-\tau}) \ .$$

Suppose that $q(0) = 0$ is satisfied, then one can estimate β by the following fixed effects estimator:

$$\widetilde{\beta} = \arg\min_{q(0)'\beta = 0} \sum_{i=1}^{N}\sum_{t=1}^{T}\left[y_{it} - q(x_{it})'\beta - \frac{1}{T}\sum_{s=1}^{T}(y_{is} - q(x_{is})'\beta)\right]^2 .$$

$g(x_{it})$ is estimated by $\widetilde{g}(x_{it}) = q(x_{it})'\widetilde{\beta}$. The test statistic in this case is

$$\widehat{\lambda} = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\begin{array}{c} y_{it} - v_0(x_{it}, \widehat{\theta}) - \sum_{j=1}^{m} p_j(v_j(x_{it}, \widehat{\theta}))'\widehat{\pi}_j - \\ \frac{1}{T}\sum_{s=1}^{T}(y_{is} - v_0(x_{is}, \widehat{\theta}) - \sum_{j=1}^{m} p_j(v_j(x_{is}, \widehat{\theta}))'\widehat{\pi}_j) \end{array}\right] *$$
$$\left[\begin{array}{c} q(x_{it})'\widetilde{\beta} - v_0(x_{it}, \widehat{\theta}) - \sum_{j=1}^{m} p_j(v_j(x_{it}, \widehat{\theta}))'\widehat{\pi}_j - \\ \frac{1}{T}\sum_{s=1}^{T}(q(x_{is})'\widetilde{\beta} - v_0(x_{is}, \widehat{\theta}) - \sum_{j=1}^{m} p_j(v_j(x_{is}, \widehat{\theta}))'\widehat{\pi}_j) \end{array}\right] .$$

Again, the asymptotic distribution of the test statistic can be established with the techniques developed in Hong and White (1995).

One drawback of the fixed effects estimation is that it cannot estimate the effect of time-invariant regressors. In some applications, researchers may want to estimate the effect of the time-invariant regressors. To do so, researchers may impose the restrictions $E(\alpha_i|x_i, w_i) = w_i'\phi_o$ and $E(u_i|x_i, w_i, \alpha_i) = 0$, and then apply the random effects estimation. The random effects estimator is consistent if both restrictions are satisfied. The random effects estimator is inconsistent, however, if either or both restrictions are not satisfied; particularly if $E(\alpha_i|x_i, w_i) = w_i'\phi_o$ is not satisfied. Thus, to avoid potential specification errors like this, it is better to leave the conditional mean $E(\alpha_i|x_i, w_i) = h_{m+1,o}(x_i, w_i)$ unspecified and consider the following model

$$y_{it} = v_0(x_{it}, \theta_o) + \sum_{j=1}^{m} h_{jo}(v_j(x_{it}, \theta_o)) + h_{m+1,o}(x_i, w_i) + (u_{it} + \varepsilon_i) \tag{14.7}$$

where $\varepsilon_i = \alpha_i - h_{m+1,o}(x_i, w_i)$. The unknown function $h_{m+1,o}(x_i, w_i)$ can be estimated exactly the same way as other unknown functions. Let

$$p_{m+1}(x,w) = (p_{m+1,1}(x,w), p_{m+1,2}(x,w), \ldots, p_{m+1,k_o}(x,w))'$$

denote known basis functions that for some $k_{m+1} \times 1$ vector π_{m+1} and some constant $\tau_{m+1} > 0$ satisfying

$$h_{m+1,o}(x,w) = p_{m+1}(x,w)'\pi_{m+1} + O(k_{m+1}^{-\tau_{m+1}}) .$$

Suppose that the following stronger condition is satisfied:

Assumption 1' (Strictly exogenous regressors). $E(u_i|x_i,w_i) = 0$.
Replace the unknown functions with their approximations. Under Assumption 1', the random effects estimator is given by

$$(\widetilde{\theta},\widetilde{\pi},\widetilde{\pi}_{m+1}) = \arg \min_{\theta,\pi_j,1\le j\le m+1} \sum_{i=1}^{N}\sum_{t=1}^{T} \left[\begin{array}{c} y_{it} - v_0(x_{it},\theta) - \\ \sum_{j=1}^{m} p_j(v_j(x_{it},\theta))'\pi_j - p_{m+1}(x_i,w_i)'\pi_{m+1} \end{array} \right]^2 .$$

The unknown functions $h_j(\cdot)$ for $j = 1,\ldots,m+1$ are estimated by $\widetilde{h}_j(\cdot) = p_j(\cdot)'\widetilde{\pi}_j$. Again, under conditions similar to those of Shen (1997) and Newey (1997), it can be shown that $\widetilde{\theta}$ is root-N consistent and asymptotically normally distributed, and any smooth functional of $\widetilde{h}_j(\cdot)$ for $j = 1,\ldots,m+1$, is asymptotically normally distributed. The asymptotic variance-covariance of $\widetilde{\theta}$ can be estimated consistently by the usual random effects covariance matrix of $\widetilde{\theta}$. Denote $\gamma = (\theta',\pi',\pi'_{m+1})'$, $\widetilde{\gamma} = (\widetilde{\theta}',\widetilde{\pi}',\widetilde{\pi}'_{m+1})'$ and denote

$$l_2(y_i,x_i,\gamma) = \sum_{t=1}^{T} \left[y_{it} - v_0(x_{it},\theta) - \sum_{j=1}^{m} p_j(v_j(x_{it},\theta))'\pi_j - p_{m+1}(x_i,w_i)'\pi_{m+1} \right]^2 .$$

Denote

$$\begin{aligned} \widetilde{V} &= \left(\sum_{i=1}^{N} \frac{\partial^2 l_2(y_i,x_i,\widetilde{\gamma})}{\partial\gamma\partial\gamma'} \right)^{-1} \times \left(\sum_{i=1}^{N} \frac{\partial l_2(y_i,x_i,\widetilde{\gamma})}{\partial\gamma} \frac{\partial l_2(y_i,x_i,\widetilde{\gamma})}{\partial\gamma'} \right) \\ &\quad \times \left(\sum_{i=1}^{N} \frac{\partial^2 l_2(y_i,x_i,\widetilde{\gamma})}{\partial\gamma\partial\gamma'} \right)^{-1} \\ &= \begin{pmatrix} \widetilde{V}_\theta & \widetilde{V}_{\theta 1} \\ \widetilde{V}_{1\theta} & \widetilde{V}_{11} \end{pmatrix} . \end{aligned}$$

Then, $\widetilde{V}$ is the usual random effects covariance matrix of $\widetilde{\gamma}$ and the upper-left block $\widetilde{V}_\theta$ is the covariance matrix of $\widetilde{\theta}$.

Notice that $\widetilde{h}_{m+1}(x_i,w_i)$ estimates the part of the individual effect that is correlated with the explanatory variables. It can be used for testing specification such as $E(\alpha_i|x_i,w_i) = w_i'\phi_o$. In principal, a test statistic based on

$$\sum_{i=1}^{N} \left(\widetilde{h}_{m+1}(x_i,w_i) - w_i'\widetilde{\phi} \right)^2 ,$$

where $\widetilde{\phi}$ is the random effects estimator of the coefficients on the time-invariant regressors, can be constructed with techniques similar to those developed by Hong and White (1995). The estimator $\widetilde{h}_{m+1}(x_i, w_i)$, however, cannot be used for testing the presence of individual effect. A kernel-based test has been proposed by Li and Wang (1998) and implemented by Kniesner and Li (2002).

One potential criticism of model (14.7) is that the unknown function h_{m+1} depends on too many regressors. Unless researchers have a very large cross sectional sample, practically this model cannot be estimated with high precision. An alternative approach is to write $E(\alpha_i|w_i) = h_{oo}(w_i)$. With $\eta_i = \alpha_i - h_{oo}(w_i)$, the model now becomes

$$y_{it} = v_0(x_{it}, \theta_o) + \sum_{j=1}^{m} h_j(v_j(x_{it}, \theta_o)) + h_{oo}(w_i) + (u_{it} + \eta_i) \,. \tag{14.8}$$

Now let $p_o(w) = (p_{o1}(w), p_{o2}(w), \ldots, p_{ok_o}(w))'$ denote known basis functions that for some $k_o \times 1$ vector π_{o1} and some scalar $\tau_o > 0$ satisfies

$$h_{oo}(w) = p_o(w)'\pi_o + O(k_o^{-\tau_o}) \,.$$

Recall that $l_1(y_i, x_i, \delta)$ is given in (14.5). Denote

$$l_3(y_i, x_i, \delta, \pi_o) = \sum_{t=1}^{T} \left[y_{it} - v_0(x_{it}, \theta) - \sum_{j=1}^{m} p_j(v_j(x_{it}, \theta))'\pi_j - p_{o1}(w_i)'\pi_o \right]^2 .$$

The parameter θ and the unknown coefficient vector π_o can be estimated jointly by solving the following equations:

$$\sum_{i=1}^{N} \frac{\partial l_1(y_i, x_i, \overline{\theta}, \overline{\pi})}{\partial \delta} = 0 \,,$$

$$\sum_{i=1}^{N} \frac{\partial l_3(y_i, x_i, \overline{\theta}, \overline{\pi}, \overline{\pi}_o)}{\partial \pi_o} = 0 \,.$$

In other words, $(\overline{\theta}, \overline{\pi}, \overline{\pi}_o)$ is the moment estimator solving the above moment conditions. The unknown functions are estimated by $\overline{h}_j(\cdot) = p_j(\cdot)'\overline{\pi}_j, j = 0, 1, \ldots, m$. Under some sufficient conditions, it can be shown that the estimator $\overline{\theta}$ and $\overline{h}_j(\cdot)$ are consistent and that $\overline{\theta}$ is asymptotically normally distributed and its asymptotic variance–covariance can be estimated consistently by the usual covaraince matrix of $\overline{\theta}$, a corresponding submatrix of the usual covariance matrix of the moment estimator $(\overline{\theta}, \overline{\pi}, \overline{\pi}_o)$.

The strict exogeneity condition rules out endogenous regressors and predetermined regressors such as the lagged dependent variables. This condition, however, can be relaxed if instrumental variables are available. To demonstrate, consider model (14.2) again. Now eliminating the individual effect by simple time-differencing, we obtain

$$y_{it} - y_{is} = v_0(x_{it}, \theta_o) + \sum_{j=1}^{m} h_{jo}(v_j(x_{it}, \theta_o)) - [v_0(x_{is}, \theta_o) + \sum_{j=1}^{m} h_{jo}(v_j(x_{is}, \theta_o))] + u_{it} - u_{is} .$$

Suppose that there exists the variables z_{its} that satisfy

Assumption 2 (instruments). *For all i and all* $s \neq t$, $E(u_{it} - u_{is}|z_{its}) = 0$.

Suppose that Assumption 2 uniquely identifies the parameter of interest. Again, we will replace the unknown functions with their approximations. In addition, we will approximate the conditional expectation by finite number of unconditional expectations. Specifically, let $r_{st}(z_{its}) = (r_{st1}(z_{its}), \ldots, r_{stk_{st}}(z_{its}))'$ denote known basis functions that approximate any measurable and square integrable function of z_{its}. With $h = (h_1(\cdot), \ldots, h_m(\cdot))$, define

$$\rho_{it}(\theta, h) = y_{it} - v_0(x_{it}, \theta) - \sum_{j=1}^{m} h_j(v_j(x_{it}, \theta)) .$$

Assumption 2 implies

$$E\{[\rho_{it}(\theta_o, h_o) - \rho_{is}(\theta_o, h_o)] \times r_{st}(z_{its})\} = 0 \text{ for any } s \neq t .$$

The above unconditional moment conditions do not uniquely identify the parameter (θ_o, h_o) because h_o is infinite dimensional. But, when these unknown functions are replaced by their approximations, we will assume that the following moment conditions

$$E\{(\rho_{it}(\theta_o, h_k) - \rho_{is}(\theta_o, h_k)) \times r_{st}(z_{its})\} = 0 \text{ for any } s \neq t ,$$

with $h_k = (p_1(\cdot)'\pi_1, \ldots, p_m(\cdot)'\pi_m)$, uniquely identifies (θ_o, π). This identification requires that the number of moment conditions $\sum_{s \neq t} k_{st}$ is larger than the number of coefficients to be estimated, $k + d_\theta$. Let $R_{1i}(\theta, \pi)$ denote the column vector formed by

$$(\rho_{it}(\theta_o, h_k) - \rho_{is}(\theta_o, h_k)) \times r_{st}(z_{its}) \text{ for all } s \neq t .$$

Then, $E\{R_{1i}(\theta, \pi)\} = 0$. The Generalized Method of Moments (hereafter GMM) estimator based on these moment restrictions is given by

$$(\widehat{\theta}^{IV}, \widehat{\pi}^{IV}) = \arg\min_{\theta, \pi} \left(\sum_{i=1}^{N} R_{1i}(\theta, \pi) \right)' \left(\sum_{i=1}^{N} R_{1i}(\theta, \pi) \right) .$$

The unknown functions are estimated by $\widehat{h}_j^{IV} = p_j(\cdot)'\widehat{\pi}_j^{IV}$ for all j. Ai and Chen (2003, 2005) show that $\widehat{\theta}^{IV}$ is root-N consistent and asymptotically normally distributed and its asymptotic variance-covariance matrix is estimated consistently by the usual GMM covariance matrix of $\widehat{\theta}^{IV}$. Recall that $\delta = (\theta', \pi')'$. Denote

$$\begin{aligned}\widehat{V}^{IV} &= \left[\left(\sum_{i=1}^{N}\frac{\partial R_{1i}(\widehat{\theta}^{IV},\widehat{\pi}^{IV})}{\partial\delta'}\right)'\left(\sum_{i=1}^{N}\frac{\partial R_{1i}(\widehat{\theta}^{IV},\widehat{\pi}^{IV})}{\partial\delta'}\right)\right]^{-1}\\ &\quad\times\left(\sum_{i=1}^{N}R_{1i}(\widehat{\theta}^{IV},\widehat{\pi}^{IV})\times R_{1i}(\widehat{\theta}^{IV},\widehat{\pi}^{IV})'\right)\\ &\quad\times\left[\left(\sum_{i=1}^{N}\frac{\partial R_{1i}(\widehat{\theta}^{IV},\widehat{\pi}^{IV})}{\partial\delta'}\right)'\left(\sum_{i=1}^{N}\frac{\partial R_{1i}(\widehat{\theta}^{IV},\widehat{\pi}^{IV})}{\partial\delta'}\right)\right]^{-1}\\ &= \begin{pmatrix}\widehat{V}_{\theta}^{IV} & \widehat{V}_{\theta\pi}^{IV}\\ \widehat{V}_{\pi\theta}^{IV} & \widehat{V}_{\pi}^{IV}\end{pmatrix}.\end{aligned}$$

Then, $\widehat{V}^{IV}$ is the usual covariance matrix of the GMM estimator $\widehat{\delta}$. The upper-left block $\widehat{V}_{\theta}^{IV}$ is the covariance matrix of $\widehat{\theta}^{IV}$.

For model (1.7), we assume that the instrumental variables z_{it} satisfies the following assumption.

Assumption 3 (instruments). *For all i and t, $E(\varepsilon_i + u_{it}|z_{it}) = 0$.*
Let $r_t(z_{it}) = (r_{t1}(z_{it}),\ldots,r_{tk_t}(z_{it}))'$ denote known basis functions that approximate any measurable and square integrable function. Suppose that the number of moment conditions $\sum_t k_t$ is larger than the number of coefficients $k + d_\theta + k_{m+1}$. Let $R_{2i}(\theta,\pi,\pi_{m+1})$ denote the column vector formed by

$$\left(y_{it} - v_0(x_{it},\theta_o) - \sum_{j=1}^{m} p_j(v_j(x_{it},\theta_o))'\pi_j - p_{m+1}(x_i,w_i)'\pi_{m+1}\right)\times r_t(z_{it}) \text{ for all } t\,.$$

Then, $E\{R_{2i}(\theta,\pi)\} = 0$. The GMM estimator based on these moment restrictions is given by

$$(\widetilde{\theta}^{IV},\widetilde{\pi}^{IV},\widetilde{\pi}_{m+1}^{IV}) = \arg\min_{\theta,\pi,\pi_{m+1}}\left(\sum_{i=1}^{N}R_{2i}(\theta,\pi,\pi_{m+1})\right)'\left(\sum_{i=1}^{N}R_{2i}(\theta,\pi,\pi_{m+1})\right).$$

The unknown functions are estimated by $\widetilde{h}_j^{IV} = p_j(\cdot)'\widetilde{\pi}_j^{IV}$ for all j. Again, Ai and Chen (2003, 2005) show that $\widetilde{\theta}^{IV}$ is root-N consistent and asymptotically normally distributed and its asymptotic variance–covariance is estimated consistently by the usual GMM covariance matrix of $\widetilde{\theta}^{IV}$. Similar estimators also can be constructed for model (14.8). See Das (2003) for further extensions.

14.2.2 Multiplicative Effect

The additive structure of the model above, though common in empirical work, is restrictive in the sense that the model assumes that the marginal effects of the explanatory variables are identical for all individuals with the same explanatory

variable values. This restriction can be relaxed with a multiplicative individual effect such as in the following model

$$y_{it} = v_0(x_{it}, \theta_o) + \sum_{j=1}^{m-1} h_j(v_j(x_{it}, \theta_o)) + \alpha_i h_m(v_m(x_{it}, \theta_o)) + u_{it} \ .$$

In this model, the marginal effects of the explanatory variables vary with individuals via the unknown function $h_m(\cdot)$. We now illustrate how the estimators discussed above are extended to models of this sort.

For identification purpose, the function $h_m(\cdot)$ is assumed to satisfy the normalization: $h_m(0) = 1$. Because the individual effect is multiplicative, simple time-differencing will not eliminate the individual effect. However, it is still possible to solve for the individual effect from one period (t) and then substitute it into another period $(s < t)$ to obtain:

$$\left\{ y_{it} - v_0(x_{it}, \theta_o) - \sum_{j=1}^{m-1} h_j(v_j(x_{it}, \theta_o)) - u_{it} \right\} h_m(v_m(x_{is}, \theta_o))$$
$$= \left\{ y_{is} - v_0(x_{is}, \theta_o) - \sum_{j=1}^{m-1} h_j(v_j(x_{is}, \theta_o)) - u_{is} \right\} h_m(v_m(x_{it}, \theta_o)) \ .$$

Define

$$\rho_{its}(\theta, h) = \left[y_{it} - v_0(x_{it}, \theta) - \sum_{j=1}^{m-1} h_j(v_j(x_{it}, \theta)) \right] h_m(v_m(x_{is}, \theta)) \ .$$

Suppose that

$$E\{\rho_{its}(\theta_o, h_0) - \rho_{ist}(\theta_o, h_o) | z_{its}\} = 0 \text{ for all } s \neq t \ .$$

Let $R_{3i}(\theta, \pi)$ denote the column vector formed by

$$(\rho_{its}(\theta_o, h_k) - \rho_{ist}(\theta_o, h_k)) \times r_{st}(z_{its}) \text{ for all } s \neq t \ .$$

Then, $E\{R_{3i}(\theta, \pi)\} = 0$. The GMM estimator based on these moment restrictions is given by

$$(\widehat{\theta}^{IV}, \widehat{\pi}^{IV}) = \arg\min_{\theta, \pi} \left(\sum_{i=1}^{N} R_{3i}(\theta, \pi) \right)' \left(\sum_{i=1}^{N} R_{3i}(\theta, \pi) \right) \ .$$

The unknown functions are estimated by $\widehat{h}_j^{IV} = p_j(\cdot)' \widehat{\pi}_j^{IV}$ for all j. It follows from Ai and Chen (2003, 2005) that $\widehat{\theta}^{IV}$ is root-N consistent and asymptotically normally distributed and its asymptotic variance–covariance is estimated consistently by the usual GMM covariance matrix of $\widehat{\theta}^{IV}$.

The marginal effect of the time-invariant regressors can be estimated by simple least squares. Substituting for the individual effect $\alpha_i = h_o(x_i, w_i) + \varepsilon_i$ yields:

$$y_{it} = v_0(x_{it},\theta_o) + \sum_{j=1}^{m-1} h_j(v_j(x_{it},\theta_o)) + h_o(x_i,w_i)h_m(v_m(x_{it},\theta_o)) + \varepsilon_i h_m(v_m(x_{it},\theta_o)) + u_{it}\ .$$

Suppose that $E\{u_{it}|x_i,w_i\} = 0$ is satisfied. Then the coefficients θ_o and $\pi = (\pi_0',\ldots,\pi_m')'$ can be estimated by the following nonlinear least squares:

$$(\widehat{\theta},\widehat{\pi}) = \arg\min \sum_{i=1}^{N}\sum_{t<s} \left\{ \begin{array}{c} y_{it} - v_0(x_{it},\theta) - \sum_{j=1}^{m-1} q_j(v_j(x_{it},\theta))'\pi_j \\ -[q_m(v_m(x_{is},\theta))'\pi_m]p_o(x_i,w_i)'\pi_o \end{array} \right\}^2 .$$

The asymptotic properties of the estimator can be derived by applying the results of Shen (1997).

14.3 Nonlinear Panel Data Model

A key structure of the linear panel data model, which is exploited by all of the estimators reviewed above, is that the observed dependent variable is a linear function of the individual effect. This linear relationship allows us to eliminate the individual effect through simple time-differencing. It also allows us to estimate the conditional mean $E(\alpha_i|x_i,w_i)$ function under the condition that (x_i,w_i) is mean-independent of (ε_i,u_{it}) $(\varepsilon_i = \alpha_i - E(\alpha_i|x_i,w_i))$. If the observed dependent variable is a nonlinear function of the individual effect, the individual effect cannot be eliminated through simple time-differencing and $E(\alpha_i|x_i,w_i)$ cannot be estimated consistently under the mean-independence condition. Panel data limited and qualitative dependent variable models are important examples where the observed dependent variable is a nonlinear function of the individual effect. In these models, some "nonlinear-differencing" techniques are required to remove the individual effect. The aim of the rest of this chapter is to review the "nonlinear-differencing" techniques proposed in the literature. We first review the panel data censored regression model, also known as Type I Tobit model, then the panel data discrete choice model, and lastly the panel data sample selection model, also known as Type II Tobit model.

14.3.1 Censored Regression Model

We begin with the panel data censored regression model, which is given by

$$\begin{aligned} y_{it}^* &= x_{it}'\theta_o + \alpha_i + u_{it} \\ y_{it} &= \max\{0, y_{it}^*\}, \qquad i = 1,2,\ldots,N;\ t = 1,2,\ldots,T \end{aligned} \tag{14.9}$$

In this model, the latent dependent variable y_{it}^* is linear in the individual effect α_i, but the observed dependent variable y_{it} is nonlinear in α_i, with the nonlinearity arising from censoring. A simple time-differencing of the observed dependent variable does not remove α_i. To see why simple time-differencing of the observed dependent variables does not remove the individual effect, for any period t and at the true value θ_o, write

$$y_{it} - x_{it}'\theta_o = \max\{y_{it}^* - x_{it}'\theta_o, -x_{it}'\theta_o\} = \max\{\alpha_i + u_{it}, -x_{it}'\theta_o\} .$$

Clearly, $y_{it} - x_{it}'\theta_o$ is the censored error term $\alpha_i + u_{it}$, with $-x_{it}'\theta_o$ as the censoring value. Similarly, for any period s,

$$y_{is} - x_{is}'\theta_o = \max\{\alpha_i + u_{is}, -x_{is}'\theta_o\}$$

is the censored error term $\alpha_i + u_{is}$, with $-x_{is}'\theta_o$ as the censoring value. Applying simple time-differencing, we obtain:

$$y_{it} - x_{it}'\theta_o - (y_{is} - x_{is}'\theta_o) = \max\{\alpha_i + u_{it}, -x_{it}'\theta_o\} - \max\{\alpha_i + u_{is}, -x_{is}'\theta_o\} .$$

The individual effect is clearly not eliminated by simple time-differencing.

From the point of view of estimating the unknown parameter θ_o, it is not necessary to remove the individual effect at every data point. As long as the differenced error term: $\max\{\alpha_i + u_{it}, -x_{it}'\theta_o\} - \max\{\alpha_i + u_{is}, -x_{is}'\theta_o\}$ has a zero conditional mean given the explanatory variables, the parameter θ_o can be estimated consistently by standard regression techniques. Unfortunately, the differenced error term does not have a zero conditional mean when: (i) the error terms u_{it} and u_{is}, conditional on the regressors and the individual effect, are not identically distributed or (ii) the censoring values $-x_{it}'\theta_o$ and $-x_{is}'\theta_o$ are not identical. Thus, to obtain a consistent estimator of θ_o, condition (i) and (ii) must not be satisfied. Condition (i) is ruled out by the following condition:

Assumption 4. *The error terms u_{it} and u_{is}, conditional on $(x_{it}, x_{is}, \alpha_i)$, are identically distributed.*

Condition (ii) is satisfied by artificially censoring the observed dependent variables so that both error terms are censored at the same value $\max\{-x_{it}'\theta_o, -x_{is}'\theta_o\}$. Specifically, define the artificially censored error terms as:

$$\begin{aligned} e(y_{it} - x_{it}'\theta_o, x_{is}'\theta_o) &= \max\{y_{it} - x_{it}'\theta_o, -x_{is}'\theta_o\} \\ &= \max\{\alpha_i + u_{it}, -x_{it}'\theta_o, -x_{is}'\theta_o\} , \end{aligned}$$

$$\begin{aligned} e(y_{is} - x_{is}'\theta_o, x_{it}'\theta_o) &= \max\{y_{is} - x_{is}'\theta_o, -x_{it}'\theta_o\} \\ &= \max\{\alpha_i + u_{is}, -x_{it}'\theta_o, -x_{is}'\theta_o\} . \end{aligned}$$

It then follows from Assumption 4 that $e(y_{it} - x_{it}'\theta_o, x_{is}'\theta_o)$ and $e(y_{is} - x_{is}'\theta_o, x_{it}'\theta_o)$, conditional on the explanatory variables and the individual effect, are identically distributed. This in turn implies that

$$E\{e(y_{it}-x_{it}'\theta_o, x_{is}'\theta_o) - e(y_{is}-x_{is}'\theta_o, x_{it}'\theta_o)|x_{it}, x_{is}\} = 0\,. \tag{14.10}$$

θ_o now can be estimated consistently from the conditional moment restriction (14.10) by standard regression techniques such as GMM, provided some identification condition is satisfied. Since (14.10) is obtained by applying simple time-differencing after some nonlinear transformation of the observed dependent variable, this approach is called "nonlinear-differencing".

Although the true value θ_o can be estimated consistently from (14.10) by GMM, estimation methods using zero conditional mean conditions are often more complex than the methods using zero unconditional mean conditions. A simpler and better approach is to find a convex objective function whose first order condition coincides with some unconditional moment conditions implied by (14.10). For instance, the objective function

$$A(\theta_0) \stackrel{\text{def}}{=} E\{r(y_{it}, y_{is}, (x_{it}-x_{is})'\theta_o)\}\,,$$

with

$$r(y_1, y_2, \delta) = \begin{cases} y_1^2/2 - \delta y_1 - y_1 y_2 & \text{if } \delta \le -y_2\,; \\ (y_1 - y_2 - \delta)^2/2 & \text{if } -y_2 < \delta < y_1\,; \\ y_2^2/2 + \delta y_2 - y_1 y_2 & \text{if } y_1 \le \delta_i\,; \end{cases}$$

satisfies the following unconditional moment condition

$$\frac{\partial A(\theta)}{\partial \theta}|_{\theta=\theta_0} = E\left\{(e(y_{it}-x_{it}'\theta_o, x_{is}'\theta_o) - e(y_{is}-x_{is}'\theta_o, x_{it}'\theta_o))(x_{it}-x_{is})\right\} = 0\,,$$

which is obviously implied by (14.10). Given that $r(y_1, y_2, \delta) \ge 0$ for all (y_1, y_2, δ), this suggests a nonlinear least squares estimator of θ_o:

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N} \sum_{t<s} r\left(y_{it}, y_{is}, (x_{it}-x_{is})'\theta\right)\,. \tag{14.11}$$

Under some sufficient conditions, Honoré (1992) shows that $\widehat{\theta}$ is consistent and asymptotically normally distributed and its asymptotic variance-covariance is estimated consistently by the usual nonlinear least squares covariance matrix of $\widehat{\theta}$:

$$\begin{aligned} \widehat{V}_\theta = & \left(\sum_{i=1}^{N} \sum_{t<s} \frac{\partial^2 r\left(y_{it}, y_{is}, (x_{it}-x_{is})'\widehat{\theta}\right)}{\partial \delta^2} (x_{it}-x_{is})(x_{it}-x_{is})' \right)^{-1} \\ & \times \left(\begin{array}{c} \sum_{i=1}^{N} \left[\sum_{t<s} \frac{\partial r\left(y_{it}, y_{is}, (x_{it}-x_{is})'\widehat{\theta}\right)}{\partial \delta} (x_{it}-x_{is}) \right] \\ \times \left[\sum_{t<s} \frac{\partial r\left(y_{it}, y_{is}, (x_{it}-x_{is})'\widehat{\theta}\right)}{\partial \delta} (x_{it}-x_{is})' \right] \end{array} \right) \\ & \times \left(\sum_{i=1}^{N} \sum_{t<s} \frac{\partial^2 r\left(y_{it}, y_{is}, (x_{it}-x_{is})'\widehat{\theta}\right)}{\partial \delta^2} (x_{it}-x_{is})(x_{it}-x_{is})' \right)^{-1}\,. \end{aligned}$$

Notice that Assumption 4 implies that

$$E\left\{\xi(e(y_{it}-x_{it}'\theta_o,x_{is}'\theta_o))-\xi(e(y_{is}-x_{is}'\theta_o,x_{it}'\theta_o))|x_{it},x_{is}\right\}=0 \tag{14.12}$$

holds for any function $\xi(\cdot)$. Honoré's (1992) estimator clearly does not use all information. More efficient estimator can be constructed from (14.12). For some integers k_1 and k_2, let $q(u)=(q_1(u),q_2(u),\ldots,q_{k_1}(u))'$ denote known basis functions that approximate any square integrable function of u, and let $p(x_{it},x_{is})=(p_1(x_{it},x_{is}),p_2(x_{it},x_{is}),\ldots,p_{k_2}(x_{it},x_{is}))'$ denote known basis functions that approximate any square integrable function of (x_{it},x_{is}). Condition (14.12) implies

$$E\{(q(e(y_{it}-x_{it}'\theta_o,x_{is}'\theta_o))-q(e(y_{is}-x_{is}'\theta_o,x_{it}'\theta_o)))\otimes p(x_{it},x_{is})\}=0,\ \text{for } t>s$$

where $\otimes$ denotes the Kronecker product. Denote

$$\rho(y_i,x_i,\theta)=\operatorname{vec}\left\{\begin{array}{c}[q(e(y_{it}-x_{it}'\theta,x_{is}'\theta))-q(e(y_{is}-x_{is}'\theta,x_{it}'\theta))]\otimes p(x_{it},x_{is}),\\ t=s+1,\ldots,T;\ s=1,2,\ldots,T-1\end{array}\right\}.$$

The unknown parameter θ_o can be estimated by GMM:

$$\widehat{\theta}^{\mathrm{GMM}}=\arg\min_{\theta}\left(\sum_{i=1}^{N}\rho(y_i,x_i,\theta)\right)'\widehat{\Omega}^{-1}\left(\sum_{i=1}^{N}\rho(y_i,x_i,\vartheta)\right),$$

where $\widehat{\Omega}$ is some known positive definite matrix, or by the empirical likelihood method:

$$\widehat{\theta}^{\mathrm{EL}}=\arg\min_{\theta}\max_{\lambda}\sum_{i=1}^{N}\ln(1+\lambda'\rho(y_i,x_i,\theta)).$$

By allowing k_1 and k_2 to grow with sample size, Ai (2005) shows that $\widehat{\theta}^{\mathrm{GMM}}$ is consistent and asymptotically normally distributed and that it is efficient for model (14.12). The asymptotic variance–covariance of $\widehat{\theta}^{\mathrm{GMM}}$ is estimated consistently by the usual GMM covariance matrix of $\widehat{\theta}^{\mathrm{GMM}}$. The asymptotic distribution of $\widehat{\theta}^{\mathrm{EL}}$ also can be derived with the techniques developed in Donald, Imbens, and Newey (2004). A drawback of the GMM estimation is that the objective function may not be globally convex and may have many local minimizers. But this problem can be resolved with Honoré's (1992) estimator as the starting value.

Assumption 4 can be strengthened by requiring the error terms to satisfy the conditional pairwise exchangeability condition:

Assumption 4'. *The pair* (u_{it},u_{is}) *is identically distributed as* (u_{is},u_{it}) *conditional on* (x_{it},x_{is},α_i).

This condition implies that

$$E\left\{\xi\left[e(y_{it}-x_{it}'\theta_o,x_{is}'\theta_o)\right]-\xi\left[e(y_{is}-x_{is}'\theta_o,x_{it}'\theta_o)\right]|x_{it},x_{is}\right\}=0 \tag{14.13}$$

for any odd function $\xi(\cdot)$. Since $\xi(u)=u$ is an odd function, Honoré's (1992) estimator is still consistent under Assumption 4'. Other consistent estimators can

be constructed analogously for arbitrary odd function. See Honoré and Kyriazidou (2000) and Charlier et al. (2000) for details. Efficient GMM estimator for this model can also be constructed in the same manner as for (14.12) by requiring that $q(u)$ are odd functions.

Under Assumption 4, Honoré's (1992) estimator can be easily extended to the following partially additive panel data Tobit model:

$$y_{it} = \begin{cases} x_{0it}\theta_o + \sum_{j=1}^{m} h_j(x_{jit}) + \alpha_i + u_{it}, & \text{if the RHS} > 0; \text{ and} \\ 0 & \text{otherwise,} \end{cases}$$

where $h_j(\cdot)$ are unknown functions. For identification purpose, we assume that the unknown functions satisfy the location restriction $h_j(0) = 0$ for all j and the exclusive restriction that $x_{1it}, \ldots, x_{mit}$ are mutually exclusive. Suppose that each $h_j(\cdot)$ is approximated by the linear sieve $p_j^{k_j}(\cdot)'\pi_j$, where $p_j^{k_j}(\cdot)$ is a vector of approximating functions satisfying $p_j^{k_j}(0) = 0$. The unknown parameter θ_o and the coefficients $\pi = (\pi_1', \ldots, \pi_m')'$ are estimated by

$$(\widehat{\theta}, \widehat{\pi}) = \arg\min_{\theta,\pi} \sum_{i=1}^{N} \sum_{t<s} r(y_{it}, y_{is}, (x_{0it} - x_{0is})'\theta + \sum_{j=1}^{m} (p_j^{k_j}(x_{jit}) - p_j^{k_j}(x_{jis}))'\pi_j) \, .$$

The unknown functions are estimated by $\widehat{h}_j(\cdot) = p_j^{k_j}(\cdot)'\widehat{\pi}_j$. Ai and Li (2005) show that the estimator $(\widehat{\theta}, \widehat{h}_1, \ldots, \widehat{h}_m)$ is consistent and derive its asymptotic distribution. The asymptotic variance–covariance of $\widehat{\theta}$ is estimated consistently by the usual nonlinear least squares covariance matrix of $\widehat{\theta}$, which is the corresponding submatrix of the nonlinear least squares covariance matrix of $(\widehat{\theta}, \widehat{\pi})$. The usual nonlinear least squares covariance matrix of $\widehat{\pi}$, however, should not be viewed as the estimator for the asymptotic variance–covariance of $\widehat{\pi}$ because $\widehat{\pi}$ does not have a asymptotic normal distribution.

By exploiting all moment conditions, efficient GMM estimator for this model can be constructed in the same manner as for (14.12). Let x_{it} denote the union of $x_{0it}, \ldots, x_{mit}$. Denote

$$\rho(y_i, x_i, \theta, \pi) = \operatorname{vec} \left\{ \begin{array}{c} (A(y_{it}, x_{it}, x_{is}, \theta, \pi) - A(y_{is}, x_{is}, x_{it}, \theta, \pi)) \otimes p(x_{it}, x_{is}) \\ t = s+1, \ldots, T; \; s = 1, \ldots, T-1 \end{array} \right\}$$

where

$$A(y_{it}, x_{it}, x_{is}, \theta, \pi) = q\left(e\left[y_{it} - x_{0it}'\theta - \sum_{j=1}^{m} p_j^{k_j}(x_{jit})'\pi_j, x_{0is}'\theta + \sum_{j=1}^{m} p_j^{k_j}(x_{jis})'\pi_j \right] \right) \, .$$

The unknown parameter θ_o is either estimated by GMM:

$$(\widehat{\theta}^{\mathrm{GMM}}, \widehat{\pi}^{\mathrm{GMM}}) = \arg\min_{\theta,\pi} \left(\sum_{i=1}^{N} \rho(y_i, x_i, \theta, \pi) \right)' \widehat{\Omega}^{-1} \left(\sum_{i=1}^{N} \rho(y_i, x_i, \theta, \pi) \right)$$

or by empirical likelihood:

$$\widehat{\theta}^{EL} = \arg\min_{\theta,\pi}\max_{\lambda}\sum_{i=1}^{N}\ln\left(1+\lambda'\rho(y_i,x_i,\theta,\pi)\right) .$$

By allowing k_1 and k_2 to grow with sample size at certain rates, Ai and Li (2005) show that $\widehat{\theta}^{\text{GMM}}$ is consistently and asymptotically normally distributed and $\widehat{h}_j = p_j^{k_j}(\cdot)'\widehat{\pi}_j^{\text{GMM}}$ is consistent. The asymptotic variance–covariance of $\widehat{\theta}^{\text{GMM}}$ is estimated consistently by the usual GMM covariance matrix of $\widehat{\theta}^{\text{GMM}}$, which is the corresponding submatrix of the usual GMM covariance matrix of $(\widehat{\theta}^{\text{GMM}},\widehat{\pi}^{\text{GMM}})$.

Assumption 4 is weaker than the one we normally make for Tobit models. It permits dependent data and allows for dependence of the error term on the explanatory variables (e.g., heteroskedastic error). But it is still restrictive. It rules out predetermined or endogenous explanatory variables, for example. To allow for predetermined or endogenous explanatory variables, we must modify the nonlinear least squares and GMM procedures reviewed above. To illustrate, consider the case of predetermined regressors first. Denote $x_{it} = (x_{1it}', x_{2it}')'$ and decompose $\theta_o = (\theta_{1o}', \theta_{2o}')'$ accordingly. Suppose that x_{2it} is predetermined (e.g. lagged dependent variable). Replace Assumption 4 with

Assumption 5. *For any $t > s$, the error terms u_{it} and u_{is}, conditional on $(x_{1it}, x_{is}, \alpha_i)$, are identically distributed.*

This condition is obviously weaker than Assumption 4. To see why Assumption 5 permits predetermined regressors, suppose that u_{it} is independent of $x_{1i} = (x_{1i1}, x_{1i2}, \ldots, x_{1iT})$ and $x_{2i}^t = (x_{2i1}, x_{2i2}, \ldots, x_{2it})$. Then, for any $s < t$, (u_{it}, u_{is}) is independent of (x_{1i}, x_{2i}^s) and Assumption 5 is satisfied as long as u_{it} and u_{is} are identically distributed.

The problem with the predetermined regressors is that the censoring value $x_{it}'\theta_o$ is correlated with u_{is}. If both censoring values are used to censor both error terms, the censored error terms will not have the same distribution. One way to resolve this difficulty is to drop the predetermined regressors from $x_{it}'\theta_o$. Suppose that $x_{2it}'\theta_{2o} \geq 0$ holds with probability one. Define:

$$\begin{aligned}
\varepsilon(y_{it} - x_{it}'\theta_o, x_{1it}'\theta_{1o}, x_{is}'\theta_o) &= \max\{y_{it} - x_{it}'\theta_o, -x_{1it}'\theta_{1o}, -x_{is}'\theta_o\} \\
&= \max\{\alpha_i + u_{it}, -x_{1it}'\theta_{1o}, -x_{is}'\theta_o\}; \\
\varepsilon(y_{is} - x_{is}'\theta_o, x_{1it}'\theta_{1o}) &= \max\{y_{is} - x_{is}'\theta_o, -x_{1it}'\theta_{1o}\} \\
&= \max\{\alpha_i + u_{is}, -x_{1it}'\theta_{1o}, -x_{is}'\theta_o\} .
\end{aligned}$$

Assumption 5 and the condition $x_{2it}'\theta_{2o} \geq 0$ with probability one imply $\varepsilon(y_{it} - x_{it}'\theta_o, x_{1it}'\theta_{1o}, x_{is}'\theta_o)$ and $\varepsilon(y_{is} - x_{is}'\theta_o, x_{1it}'\theta_{1o})$ are identically distributed given (x_{1it}, x_{is}). This in turn implies:

$$E\left\{\xi(\varepsilon(y_{it} - x_{it}'\theta_o, x_{1it}'\theta_{1o}, x_{is}'\theta_o)) - \xi(\varepsilon(y_{is} - x_{is}'\theta_o, x_{1it}'\theta_{1o}))|x_{1it}, x_{is}\right\} = 0 \quad (14.14)$$

for any function $\xi(\cdot)$. The parameter θ_o now can be estimated from the conditional moment condition (14.14) by GMM or the empirical likelihood. Specifically, let $p(x_{1it}, x_{is}) = (p_1(x_{1it}, x_{is}), p_2(x_{1it}, x_{is}), \ldots, p_{k_2}(x_{1it}, x_{is}))'$ denote known

basis functions that approximate any square integrable function of (x_{it}, x_{is}). Condition (14.12) implies

$$E\{(q(\varepsilon(y_{it} - x_{it}'\theta_o, x_{1it}'\theta_{1o}, x_{is}'\theta_o)) - q(\varepsilon(y_{is} - x_{is}'\theta_o, x_{1it}'\theta_{1o}))) \otimes p(x_{1it}, x_{is})\} = 0\,, \quad \text{for } t > s$$

Denote

$$\rho(y_i, x_i, \theta) = \text{vec}\left\{ \begin{array}{c} [q(\varepsilon(y_{it} - x_{it}'\theta_o, x_{1it}'\theta_{1o}, x_{is}'\theta_o)) - q(\varepsilon(y_{is} - x_{is}'\theta_o, x_{1it}'\theta_{1o}))] \\ \otimes p(x_{1it}, x_{is}), \\ t = s+1, \ldots, T;\ s = 1, 2, \ldots, T-1 \end{array} \right\}.$$

The unknown parameter θ_o is either estimated by GMM:

$$\widehat{\theta}^{\text{GMM}} = \arg\min_{\theta} \left(\sum_{i=1}^{N} \rho(y_i, x_i, \theta) \right)' \widehat{\Omega}^{-1} \left(\sum_{i=1}^{N} \rho(y_i, x_i, \theta) \right),$$

where $\widehat{\Omega}$ is some known positive definite matrix, or by empirical likelihood:

$$\widehat{\theta}^{\text{EL}} = \arg\min_{\theta} \max_{\lambda} \sum_{i=1}^{N} \ln(1 + \lambda' \rho(y_i, x_i, \theta))\,.$$

The asymptotic distribution of $\widehat{\theta}^{\text{GMM}}$ can be derived exactly the same way as in Honoré and Hu (2004), while the asymptotic distribution of $\widehat{\theta}^{\text{EL}}$ can be derived exactly the same way as in Donald, Imbens, and Newey (2004).

The condition that $x_{2it}'\theta_{2o} \geq 0$ holds with probability one is critical for the above estimator. This condition appears very restrictive. Fortunately in most applications, the predetermined regressors are lagged dependent variables that are always non-negative and usually have positive coefficients. In those applications, this condition is imposed through restriction on the coefficients $\theta_2 \geq 0$.

Next, consider the case where (x_{2it}, x_{2is}) are endogenous. Let (z_{it}, z_{is}) denote the instrumental variables for (x_{2it}, x_{2is}). Depending on the restrictions we impose on the endogenous regressors, we may make one of the following two assumptions:

Assumption 6. *The two terms* $(u_{it}, x_{2it}, x_{2is})$ *and* $(u_{is}, x_{2it}, x_{2is})$, *conditional on*$(x_{1it}, x_{1is}, z_{it}, z_{is}, \alpha_i)$, *are identically distributed.*

Assumption 6'. *The error terms* u_{it} *and* u_{is}, *conditional on* $(x_{1it}, x_{1is}, z_{it}, z_{is}, \alpha_i)$, *are identically distributed.*

Under Assumption 6, $e(y_{it} - x_{it}'\theta_o, x_{is}'\theta_o)$ and $e(y_{is} - x_{is}'\theta_o, x_{it}'\theta_o)$, conditional on $(x_{1it}, x_{1is}, z_{it}, z_{is}, \alpha_i)$, are identically distributed, implying that

$$E\{\xi(e(y_{it} - x_{it}'\theta_o, x_{is}'\theta_o)) - \xi(e(y_{is} - x_{is}'\theta_o, x_{it}'\theta_o)) | x_{1it}, x_{1is}, z_{it}, z_{is}\} = 0 \quad (14.15)$$

for any function $\xi(\cdot)$. Thus θ_o can be estimated consistently from the conditional moment restriction (14.15) by GMM or empirical MLE. Notice that condition (14.15) does not require that $x'_{2it}\theta_{2o} \geq 0$ and $x'_{2is}\theta_{2o} \geq 0$ hold with probability one. If we require that $x'_{2it}\theta_{2o} \geq 0$ and $x'_{2is}\theta_{2o} \geq 0$ hold with probability one, we can drop x_{2it} and x_{2is} from the censoring values and modify the censored error terms as:

$$\begin{aligned}
\upsilon(y_{it} - x'_{it}\theta_o, x'_{1it}\theta_{1o}, x'_{1is}\theta_{1o}) &= \max\{y_{it} - x'_{it}\theta_o, -x'_{1it}\theta_{1o}, -x'_{1is}\theta_{1o}\} \\
&= \max\{\alpha_i + u_{it}, -x'_{1it}\theta_{1o}, -x'_{1is}\theta_{1o}\}\,; \\
\upsilon(y_{is} - x'_{is}\theta_o, x'_{1it}\theta_{1o}, x'_{1is}\theta_{1o}) &= \max\{y_{is} - x'_{is}\theta_o, -x'_{1it}\theta_{1o}, -x'_{1is}\theta_{1o}\} \\
&= \max\{\alpha_i + u_{is}, -x'_{1it}\theta_{1o}, -x'_{1is}\theta_{1o}\}\,.
\end{aligned}$$

Assumption 6' implies

$$E\left\{\begin{bmatrix} \xi\left(\upsilon(y_{it} - x'_{it}\theta_o, x'_{1it}\theta_{1o}, x'_{1is}\theta_{1o})\right) \\ -\xi\left(\upsilon(y_{is} - x'_{is}\theta_o, x'_{1it}\theta_{1o}, x'_{1is}\theta_{1o})\right) \end{bmatrix} | x_{1it}, x_{1is}, z_{it}, z_{is}\right\} = 0 \tag{14.16}$$

for any function $\xi(\cdot)$. Again, θ_o can be estimated consistently from the conditional moment restriction (14.16) by GMM or empirical MLE. The condition that $x'_{2it}\theta_{2o} \geq 0$ and $x'_{2is}\theta_{2o} \geq 0$ hold with probability one is more restrictive then previous case since the endogenous regressors may have negative coefficients.

The ideas described above can be easily extended to the following dynamic latent dependent variable model:

$$\begin{aligned}
y^*_{it} &= \rho_o y^*_{it-1} + x'_{it}\theta_o + \alpha_i + u_{it} \\
y_{it} &= \max\{0, y^*_{it}\}\,.
\end{aligned} \tag{14.17}$$

To illustrate, suppose that x_{it} is strictly exogenous. Note that this model differs from other Tobit models in that the lagged latent dependent variable may not be observed. First, we select a subsample in which $y_{it-2} > 0$ for some t. Then

$$\begin{aligned}
y_{it-1} - \rho_o y_{it-2} - x'_{it-1}\theta_o &= \max\{\alpha_i + u_{it-1}, -\rho_o y_{it-2} - x'_{it-1}\theta_o\}; \\
y_{it} - \rho_o y^*_{it-1} - x'_{it}\theta_o &= \max\{\alpha_i + u_{it}, -\rho_o y^*_{it-1} - x'_{it}\theta_o\}\,.
\end{aligned}$$

It is reasonable to assume that the variable y_{it-2} is independent of u_{it-1} and u_{it}. y^*_{it-1} (and hence y_{it-1}), on the other hand, is not independent of u_{it-1}. Assume that $\rho_o > 0$. Denote the censored error terms as

$$\begin{aligned}
&e(y_{it} - \rho_o y^*_{it-1} - x'_{it}\theta_o, \rho_o y_{it-2} + x'_{it-1}\theta_o, x'_{it}\theta_o) \\
&= \max\{y_{it} - \rho_o y^*_{it-1} - x'_{it}\theta_o, -\rho_o y_{it-2} - x'_{it-1}\theta_o, -x'_{it}\theta_o\} \\
&= \max\{\alpha_i + u_{it}, -\rho_o y_{it-2} - x'_{it-1}\theta_o, -x'_{it}\theta_o\}; \\
&e(y_{it-1} - \rho_o y_{it-2} - x'_{it-1}\theta_o, x'_{it}\theta_o) \\
&= \max\{y_{it-1} - \rho_o y_{it-2} - x'_{it-1}\theta_o, -x'_{it}\theta_o\} \\
&= \max\{\alpha_i + u_{it-1}, -\rho_o y_{it-2} - x'_{it-1}\theta_o, -x'_{it}\theta_o\}\,.
\end{aligned}$$

Although the above censored error terms are identically distributed if u_{it} and u_{it-1} are identically distributed, the first censored error term is not feasible since the lagged latent dependent variable is not observed. One way to resolve this difficulty is to impose the condition $y_{it-1} > 0$ so that the latent dependent variable is observed. However, because u_{it} is independent of, but u_{it-1} is not independent of, $y_{it-1} > 0$, the constraint $y_{it-1} > 0$ imposes restriction on u_{it-1} but not on u_{it} and, as a result, the two censored error terms are not identically distributed. To ensure that both censored error terms are identically distributed, we must impose the same constraints on u_{it-1} and u_{it}. For instance, we can require that the errors to satisfy

$$\min\{\alpha_i + u_{it-1}, \alpha_i + u_{it}\} = \max\{-\rho_o y_{it-2} - x'_{it-1}\theta_o, -x'_{it}\theta_o\} . \tag{14.18}$$

This condition implies that

$$\begin{aligned} y^*_{it-1} &\geq \max\{0, \rho_o y_{it-2} + x'_{it-1}\theta_o - x'_{it}\theta_o\} \geq 0 \text{ and hence } y_{it-1} > 0; \\ y^*_{it} &\geq \max\{\rho_o y_{it-1} + x'_{it}\theta_o - \rho_o y_{it-2} - x'_{it-1}\theta_o, \rho_o y_{it-1}\} > 0 . \end{aligned}$$

Assumption 7. *For any t, the error terms u_{it} and u_{it-1}, conditional on $(y_{it-2} > 0, x_{it}, x_{is}, \alpha_i)$, are identically distributed.*

Under assumption 7, and conditional on

$$\begin{aligned} A_{its} = \{&y_{it-2} > 0, y_{it-1} \geq \max\{0, \rho_o y_{it-2} + x'_{it-1}\theta_o - x'_{it}\theta_o\} \text{ and} \\ &y_{it} \geq \max\{\rho_o y_{it-1} + x'_{it}\theta_o - \rho_o y_{it-2} - x'_{it-1}\theta_o, \rho_o y_{it-1}\} , \end{aligned}$$

$e(y_{it} - \rho_o y_{it-1} - x'_{it}\theta_o, \rho_o y_{it-2} + x'_{it-1}\theta_o, x'_{it}\theta_o)$ and $e(y_{it-1} - \rho_o y_{it-2} - x'_{it-1}\theta_o, x'_{it}\theta_o)$ are identically distributed. This leads to the following conditional moment conditions:

$$E\left\{ 1(A_{its}) \times \begin{bmatrix} \xi(e(y_{it-1} - \rho_o y_{it-2} - x'_{it-1}\theta_o, x'_{it}\theta_o)) - \\ \xi(e(y_{it} - \rho_o y_{it-1} - x'_{it}\theta_o, \rho_o y_{it-2} + x'_{it-1}\theta_o, x'_{it}\theta_o)) \end{bmatrix} | x_{it}, x_{is} \right\} = 0$$

for any function $\xi(\cdot)$. The parameter θ_o can now be estimated from the above conditional moment restriction, and the asymptotic properties of the estimator can be derived exactly the same way as in Hu (2002).

14.3.2 Discrete Choice Model

A key aspect of the "nonlinear differencing" technique developed for the Tobit model is that the latent dependent variable is observed partly so that trimming can be used to restore the symmetry of the distribution of the observed data. This trick does not work for the panel data discrete choice model because the latent dependent variable is not observed at all. A new "nonlinear differencing" approach must be developed. In this subsection, we review those new "nonlinear differencing" techniques. We begin with the panel binary choice model given by

$$y_{it} = 1\left\{x_{it}'\theta_o + \alpha_i + u_{it} > 0\right\},\ i = 1,2,\ldots,n;\ t = 1,2,\ldots,T \tag{14.19}$$

where x_{it} is a vector of time-varying explanatory variables, α_i is an individual specific intercept, and u_{it} is the error term. Notice that, for any two time periods t and s, simple time-differencing gives:

$$y_{it} - y_{is} = 1\left\{x_{it}'\theta_o + \alpha_i + u_{it} > 0\right\} - 1\left\{x_{is}'\theta_o + \alpha_i + u_{is} > 0\right\}\ .$$

Taking expectation, we obtain:

$$\begin{aligned} &E\{y_{it} - y_{is}|x_{it}, x_{is}, \alpha_i\} \\ &= \Pr(u_{it} > -x_{it}'\theta_o - \alpha_i|x_{it}, x_{is}, \alpha_i) - \Pr(u_{is} > -x_{is}'\theta_o - \alpha_i|x_{it}, x_{is}, \alpha_i)\ . \end{aligned}$$

Obviously, simple time-differencing does not eliminate the individual effect unless $x_{it}'\theta_o = x_{is}'\theta_o$. Manski (1987), however, observes that, if the differenced probability on the right hand side has the same sign as $x_{it}'\theta_o - x_{is}'\theta_o$, then $(y_{it} - y_{is})$ is positively correlated with $sgn(x_{it}'\theta_o - x_{is}'\theta_o)$. Based on this observation, Manski proposed a maximum score estimator that maximizes the sample correlation:

$$\widehat{\theta} = \arg\max_{\theta'\theta=1} \sum_{i=1}^{n} \sum_{s<t} (y_{it} - y_{is}) * sgn((x_{it} - x_{is})'\theta)\ .$$

Clearly, Manski's estimator is defined on the differenced data through sign function. Since sign function is nonlinear, this technique is still called as "nonlinear differencing".

To ensure that the differenced probability has the same sign as $x_{it}'\theta_o - x_{is}'\theta_o$, we impose the following condition:

Assumption 8. *For any $t > s$, the error terms u_{it} and u_{is}, conditional on $(x_{it}, x_{is}, \alpha_i)$, are identically distributed.*

Like Assumption 4 for the Tobit model, Assumption 8 is weaker than the one we normally make for the binary choice model. For example, it does not require specifying the error distribution and permit dependent data and heteroskedasticity. Under some additional conditions and with some scale normalization on the parameter, Manski (1987) shows that the maximum score estimator is consistent. However, his estimator is not root-N consistent and is not asymptotically normally distributed. The nonnormal asymptotic distribution of his estimator is the result of the nonsmooth objective function. If the smoothing technique suggested by Horowitz (1992) is used here, it can be shown that the resulting estimator is asymptotically normally distributed, although the rate is still slower than root-N, (see Kyriazidou (1997) and Charlier et. al. (1995) for details).

Extension of Manski's idea to the following nonparametric panel data model

$$y_{it} = 1\left\{h(x_{it}) + \alpha_i + u_{it} > 0\right\},\ i = 1,2,\ldots,N;\ t = 1,2,\ldots,T \tag{14.20}$$

where $h(\cdot)$ is of unknown form, is straightforward. Let $p^k(x)'\pi = p_1(x)\pi_1 + \ldots + p_k(x)\pi_k$ denote the approximation to $h(x)$. Denote:

$$\widehat{\pi} = \arg\max_{\pi'\pi=1} \sum_{i=1}^{n} \sum_{s<t} (y_{it} - y_{is}) * \mathrm{sgn}((p^k(x_{it}) - p^k(x_{is}))'\pi) \, .$$

and $\widehat{h}(x) = p^k(x)'\widehat{\pi}$. Then using the techniques developed by Shen (1997) and Manski (1987), it can be shown that $\widehat{h}(x)$ is consistent under both the sup and L_2 norm.

Return to model (14.19). Like Assumptions 4, 8 rules out the predetermined explanatory variables such as the lagged dependent variables. If the predetermined explanatory variables are allowed for, the trick used by Honoré and Kyriazidou (2000) can be used here to estimate model (14.19). Specifically, decompose $x_{it}'\theta_o = x_{1it}'\theta_{1o} + x_{2it}'\theta_{2o}$. Suppose that x_{2it} are the predetermined explanatory variables. Consider three periods $r < s < t$. The insight of Honoré and Kyriazidou (2000) gives the following estimator:

$$\widehat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \sum_{r<s<t} \begin{pmatrix} K\left(\frac{x_{1is} - x_{1it}}{\sigma_n}\right) \times (y_{is} - y_{ir}) \times \\ \mathrm{sgn}((x_{1is} - x_{1ir})'\theta_1 + (x_{2it} - x_{2ir})'\theta_2) \end{pmatrix}$$

where $K(\cdot)$ denotes the kernel function and σ_n denotes the bandwidth. Again, only consistency of this estimator is proved by Honoré and Kyriazidou (2000).

For the nonparametric model (1.20) with $h(x) = h_1(x_1) + h_2(x_2)$, let $p_1^{k_1}(x_1)'\pi_1 = p_{11}(x_1)\pi_{11} + \ldots + p_{1k_1}(x_1)\pi_{1k_1}$ denote the approximation to $h_1(x_1)$ and $p_2^{k_2}(x_2)'\pi_2 = p_{21}(x_2)\pi_{21} + \ldots + p_{2k_2}(x_2)\pi_{2k_2}$ denote the approximation to $h_2(x_2)$. Then, Honoré and Kyriazidou's (2000) idea gives the following estimator:

$$\widehat{\pi} = \arg\max_{\theta} \sum_{i=1}^{n} \sum_{r<s<t} \begin{pmatrix} K\left(\frac{x_{1is} - x_{1it}}{\sigma_n}\right) \times (y_{is} - y_{ir}) \times \\ \mathrm{sgn}\left[(p_1^{k_1}(x_{1is}) - p_1^{k_1}(x_{1ir}))' \, \pi_1 + (p_2^{k_2}(x_{2it}) - p_2^{k_2}(x_{2ir}))'\pi_2\right] \end{pmatrix} .$$

Again, it can be shown that $\widehat{h}_j(x) = p_j^k(x)'\widehat{\pi}_j$ is consistent.

The important question is whether we can achieve root-N consistency and asymptotic normality. Anderson (1970) answered this question by considering the Logit version of model (14.19). His "nonlinear differencing" idea is based on a conditional maximum likelihood approach. Define $\delta_i = \sum_{t=1}^{T} y_{it}$ which takes values in $\{0, 1, \ldots, T\}$ (since $y_{it} \in \{0, 1\}$). Also define $\delta_{it} = 1$ if $y_{it} = 1$, and $\delta_{it} = 0$ otherwise. Then the conditional likelihood estimator is given by

$$\widehat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log \left(\frac{\exp(\sum_{t=1}^{T} y_{it} x_{it}'\theta)}{\sum_{\delta_{i1} + \ldots + \delta_{iT} = \delta_i} \exp(\sum_{t=1}^{T} d_{it} x_{it}'\theta)} \right) ,$$

where $\sum_{\delta_{i1} + \ldots + \delta_{iT} = \delta_i}$ denotes sum over all possible combinations of $(\delta_{i1}, \ldots, l\delta_{iT})$ with $\sum_{t=1}^{T} \delta_{it} = \delta_i$ $(= \sum_{i=1}^{T} y_{it})$. Anderson showed that the conditional maximum likelihood estimator is root-N consistent and asymptotically normally distributed.

This "nonlinear differencing" idea is easily extended to the Logit version of model (1.20), with

$$\widehat{\pi} = \arg\max_{\pi} \sum_{i=1}^{n} \log \left(\frac{\exp(\sum_{t=1}^{T} y_{it} p^k(x_{it})'\pi)}{\sum_{\delta_{i1}+\ldots+\delta_{iT}=\delta_i} \exp(\sum_{t=1}^{T} \delta_{it} p^k(x_{it})\pi)} \right)$$

and $\widehat{h}(x) = p^k(x)'\widehat{\pi}$. The consistency and the asymptotic normality of smooth functionals of $\widehat{h}(x)$ can be proved exactly as in Shen (1997).

The "nonlinear differencing" idea for the Logit model also can be extended to allow for the predetermined explanatory variables in the panel data Logit model. As showed in Honoré and Kyriazidou (2000), at least three periods of data are required; and the model parameter is estimated by:

$$\widehat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \sum_{r<s<t} 1\{y_{ir} + y_{is} = 1\} K\left(\frac{x_{1is} - x_{1it}}{\sigma_n} \right) * \log \left(\frac{[\exp((x_{1ir} - x_{1is})'\theta_1 + (x_{2ir} - x_{2it})'\theta_2)]^{y_{ir}}}{1 + \exp((x_{1ir} - x_{1is})'\theta_1 + (x_{2ir} - x_{2it})'\theta_2)} \right).$$

However, this estimator is not root-N consistent, it has the usual nonparametric kernel estimation rate of convergence.

It is clear from the above discussions that the maximum score estimator is not root-N consistent but it imposes the weakest restrictions on the distribution of the error terms. On the other hand, the conditional maximum likelihood estimator is root-N consistent but imposes the strongest assumption on the distribution of the error terms. A natural question is were there exist other restrictions on the error distribution that permit root-N consistent estimator for the model parameter. Unfortunately, Chamberlain (1993) gives a surprisingly negative answer to this question. He showed that even if the errors are i.i.d. and independent of the explanatory variables and the individual effects, the model parameter can be estimated root-N consistently only in the Logit case. Hahn (2001) considers the semiparametric information bound in dynamic panel Logit models with fixed effects. Hahn shows that the conditional maximum likelihood estimator is not semiparametrically efficient for models with only the lagged dependent variable. For more general models with regressors include time dummies, Hahn shows that the semiparametric information bound is singular, therefore, root-N consistent estimation is infeasible in more general models. Therefore, to obtain a root-N consistent estimator, it is clear that some additional assumptions must be imposed on the correlation between the explanatory variables and the individual effect. Lee (1999) takes differences across individuals in addition to the time-difference proposed by Manski (1987). Lee shows that the resulting estimator is a root-N consistent semiparametric estimator that does not depend on a smoothing parameter. Honoré and Lewbel (2002) require that there exists a "special regressor", which is continuous with bounded support and is independent of the individual effect and the error term. The role of this "special regressor" is to pull the individual effect out of the nonlinear function. Specifically,

write $x_{it}'\theta_o = x_{1it} + x_{2it}'\theta_{1o}$, where x_{1it} is that "special regressor" and its coefficient is normalized to unity for identification purpose, and x_{2it} denote the predetermined regressors. For any two periods $r > s$, let z_{is} denote the instrumental variables consisting of all predetermined variables up to time s. Under some conditions on the special regressor, Honoré and Lewbel showed that

$$E\left\{\frac{z_{is}(y_{it} - 1\{x_{1it} > 0\})}{f_t(x_{1it}|x_{2it}, z_{is})}\right\} = z_{is}x_{2it}'\theta_{1o} + E\{z_{is}\alpha_i\},\ t = r, s$$

where f_t denotes the conditional density of x_{1it} conditional on (x_{2it}, z_{is}). The individual effect now can be eliminated through simple time-differencing between period r and period s, and hence the parameter can be estimated by simple instrumental variable estimation. For details see Honoré and Lewbel (2002). Even when the parameters in a dynamic discrete choice model are not identified, it may be possible to bound the parameters in a narrow region, see Honoré and Tamer (2005) for more details on how to find the bounds for the parameters in dynamic discrete choice panel data models.

14.3.3 Sample Selection Model

The panel data sample selection model is given by:

$$\begin{aligned} y_{it} &= d_{it}(x_{it}'\theta_o + \alpha_i + u_{it}), \\ d_{it} &= 1\{z_{it}'\beta_o + \mu_i + \varepsilon_{it} > 0\}\ , \end{aligned} \tag{14.21}$$

where x_{it} and z_{it} are explanatory variables, α_i and μ_i are individual effects, and u_{it} and ε_{it} are the remainder error terms. This model consists of a binary selection equation and a regression equation which is sometimes censored. Thus, it is more complicated than the panel data binary choice model and the panel data Tobit model. The coefficient β_o can be estimated consistently by any of the methods described above for the panel discrete choice model. But to estimate the coefficient θ_o, a new differencing technique is needed. For any two periods $t > s$, note that, conditional on $d_{it} = 1, d_{is} = 1, x_{it}, z_{it}, x_{is}, z_{is}, \alpha_i, \mu_i$:

$$\begin{aligned} y_{it}|_{d_{it}=1,d_{is}=1,x_{it},z_{it},x_{is},z_{is},\alpha_i,\mu_i} &= x_{it}'\theta_o + \alpha_i + u_{it}|_{d_{it}=1,d_{is}=1,x_{it},z_{it},x_{is},z_{is},\alpha_i,\mu_i}\ , \\ y_{is}|_{d_{it}=1,d_{is}=1,x_{it},z_{it},x_{is},z_{is},\alpha_i,\mu_i} &= x_{is}'\theta_o + \alpha_i + u_{is}|_{d_{it}=1,d_{is}=1,x_{it},z_{it},x_{is},z_{is},\alpha_i,\mu_i}\ . \end{aligned}$$

If $u_{it}|_{d_{it}=1,d_{is}=1,x_{it},z_{it},x_{is},z_{is},\alpha_i,\mu_i}$ and $u_{is}|_{d_{it}=1,d_{is}=1,x_{it},z_{it},x_{is},z_{is},\alpha_i,\mu_i}$ are identically distributed, then we have

$$\begin{aligned} &E\{\xi(y_{it} - x_{it}'\theta_o)|d_{it} = 1, d_{is} = 1, x_{it}, z_{it}, x_{is}, z_{is}, \alpha_i, \mu_i\} \\ &= E\{\xi(y_{is} - x_{is}'\theta_o)|d_{it} = 1, d_{is} = 1, x_{it}, z_{it}, x_{is}, z_{is}, \alpha_i, \mu_i\}\ , \end{aligned}$$

for any function $\xi(\cdot)$. And the parameter θ_o can be estimated from the above conditional moment restriction by standard regression techniques such as GMM. The problem with this approach is that $u_{it}|_{d_{it}=1,d_{is}=1,x_{it},z_{it},x_{is},z_{is},\alpha_i,\mu_i}$ and $u_{is}|_{d_{it}=1,}$ $d_{is}=1,x_{it},z_{it},x_{is},z_{is},\alpha_i,\mu_i$ are not identically distributed unless $z_{it}'\beta_o = z_{is}'\beta_o$. This leads to the following estimator

$$\widehat{\theta} = \arg\min_{\theta} \sum_i \sum_{s<t} d_{it} d_{is} K\left(\frac{(z_{it}-z_{is})'\widehat{\beta}}{\sigma_n}\right)[(y_{it}-x_{it}'\theta)-(y_{is}-x_{is}'\theta)]^2$$

where $\widehat{\beta}$ is a consistent estimator of β, obtained by any techniques described above for the panel binary choice model. Kyiazidou (1997) proves the consistency of $\widehat{\theta}$ and derives its asymptotic distribution. Because this estimator uses a kernel weight function, the estimator is not root-N consistent, it has the standard nonparametric estimation rate of convergence.

If $u_{it}|_{d_{it}=1,d_{is}=1,x_{it},z_{it},x_{is},z_{is},\alpha_i,\mu_i}$ and $u_{is}|_{d_{it}=1,d_{is}=1,x_{it},z_{it},x_{is},z_{is},\alpha_i,\mu_i}$ satisfy the stronger exchangeability condition, then we have

$$E\{\xi(y_{it}-x_{it}'\theta_o - y_{is}+x_{is}'\theta_o)|d_{it}=1,d_{is}=1,x_{it},z_{it},x_{is},z_{is},\alpha_i,\mu_i\}=0$$

for any odd function $\xi(\cdot)$. In this case, a new class of estimator can be obtained:

$$\widehat{\theta}_{\xi} = \arg\min_{\theta} \sum_i \sum_{s<t} d_{it} d_{is} K\left(\frac{(z_{it}-z_{is})'\widehat{\beta}}{\sigma_n}\right)\Xi(y_{it}-x_{it}'\theta)-(y_{is}-x_{is}'\theta)].$$

The asymptotic distribution of the estimator $\widehat{\theta}_{\xi}$ can be derived exactly the same way as in Kyiazidou (1997).

Wooldridge (1995) also proposes some estimation methods that allow for the unobserved effects in both the regression and selection equations to be correlated with the observed variables and the error distribution in the regression equation to be unspecified. Lee (2001) proposes a semiparametric first-difference estimator for panel censored-selection models when the selection equation is of Tobit type. Also, Lee's estimator does not require smoothing.

14.4 Conclusion

In this chapter, we survey the large and growing literature on semiparametric and nonparametric panel data models. Our survey indicates that substantial progress has been made in semiparametric and nonparametric linear panel data models and panel data Tobit models. The progress made in the area of panel discrete choice and sample selection models, however, is less satisfactory, though considerable scholarly work has been devoted to this area. Far less satisfactory progress has been made in the applications of the techniques surveyed here to analyze real data. Future research in this exciting area should be focused on developing root-N

consistent estimator for the panel data discrete choice under some suitably conditions. Once the root-N consistent estimator for the panel data binary choice model is developed, the root-N consistent estimator for the panel data sample selection models should be easily constructed. Furthermore, applications of the existing techniques must be encouraged.

Due to space limitation, we do not cover topics on bias reduction techniques for nonlinear panel data models (Hahn and Newey (2004)), nor do we discuss the general nonseparable panel data models with endogenous regressors considered by Altonji and Matzkin (2005), or the panel data Poisson and duration models. See Blundell, Griffith and Windmeijer (2002), Das and Ying (2005), Horowitz and Lee (2004), Lee (2004) and Van den Berg (2001) and the references therein for discussions on nonlinear count and duration models.

Finally, our discussion on estimation of a nonparametric regression model with fixed effects is based on the within transformation (and series approximation). Baltagi and Li (2002) consider series estimation based on first-difference rather than the within transformation. Carroll, Henderson and Li (2005) propose a nonparametric kernel estimator based on first-differencing.

References

Ai, C. (2005). Some efficient estimators for panel data Tobit models, Unpublished manuscript.

Ai, C. and Chen, X. (2003). Efficient etimation of conditional moment restrictions models containing unknown functions, *Econometrica*, 71: 1795–1843.

Ai, C. and Chen, X. (2005). Asymptotic distribution of smooth functionals of sieve minimum distance estimator, Unpublished manuscript.

Ai, C. and Li, Q. (2005). Estimation of partly specified panel data Tobit models, Unpublished manuscript.

Altonji, J. and Matzkin, R. (2005). Cross section panel data estimators for nonparametric models with endogenous regressors, *Econometrica*, 73: 1053–1102.

Anderson, E. (1970). Asymptotic properties of conditional maximum likelihood estimators, *Journal of the Royal Statistical Society*, 32: 283–301.

Arellano, M. and Honore, B. (2001). Panel data models: Some recent developments, in *Handbook of Econometrics*, J. J. Heckman and E. Leamer (eds.), Vol. 5, Chap. 53, Amsterdam: North-Holland, 3229–3296.

Baltagi B. H. (2005). *Econometric Analysis of Panel Data* (3rd edition), New York: Wiley and Sons.

Baltagi, B.H. and Li, Q. (2002). Series estimation of partially linear models with fixed effects, *Annals of Economics and Finance*, 3: 103–116.

Blundell, Griffith, R. and Windmeijer, F. (2002). Individual effects and dynamics in count data models. *Journal of Econometrics*, 108: 113–131.

Carroll, R., Henderson, D. and Li, Q. (2005). Nonparametric estimation of panel data models with fixed effects, Unpublished manuscript.

Chamberlain, G. (1993). Feedback in panel data models, Unpublished manuscript.

Charlier, E., Melenberg, B. and Soest, A.V. (1995). A smoothed maximum score estimator for the binary choice panel data model and an application to labour force participation, *Statistica Neerlandica*, 49: 324–342.

Charlier, E., Melenberg, B. and Soest, A.V. (2000). Estimation of a censored regression panel data model using conditional moment restrictions efficiently, *Journal of Econometrics*, 95: 25–56.

Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data, *Econometrica*, 66: 289–314.

Chui, C.K. (1992). *An Introduction to Wavelets*. San Diego, CA: Academic Press, Inc.

Das, M. (2003). Identification and sequential estimation of panel data models with insufficient exclusion restrictions, *Journal of Econometrics*, 114: 297–328.

Das, M. and Ying, Z. (2005). Linear regression for dependently censored panel duraton models with nonadditive fixed effects, Unpublished manuscript.

Donald, S., Imbens, G. and Newey, W.K. (2004). Empirical likelihood estimation and consistent tests with conditional moment restrictions, *Journal of Econometrics*, 117: 55–93.

Fan, Y. and Li, Q. (1996). Consistent model specification tests: Omitted variable and semiparametric functional forms, *Econometrica*, 64: 865–890.

Hahn, J. (2001). The information bound of a dynamic panel Logit model with fixed effects. *Econometric Theory* 17, 913–932.

Hahn, J. and Newey, W. K. (2004). Jackknife and analytical bias reduction for nonlinear panel models, *Econometrica*, 72: 1295–1319.

Hong, Y. and White, H. (1995). Consistent specification testing via nonparametric series regression, *Econometrica,* 63: 1133–1159.

Honoré, B. (1992). Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects, *Econometrica*, 60: 533–565.

Honoré, B. and Hu, L. (2004). Estimation of cross sectional and panel data censored regression models with endogeneity, *Journal of Econometrics*, 122: 293–316.

Honoré, B. and Kyriazidou, E. (2000). Panel data discrete choice models with lagged dependent variables, *Econometrica*, 68: 839–874.

Honoré, B. and Lewbel, A. (2002). Semiparametric binary choice panel data models without strictly exogenous regressors, *Econometrica*, 70: 2053–2063.

Honoré, B. and Tamer, E. (2005). Bounds on parameters in dynamic discrete choice models, *Econometrica*, 74: 611–630.

Horowitz, J. (1992). A smoothed maximum score estimator for the binary response model, *Econometrica*, 60: 505–531.

Horowitz, J.L., and Lee, S. (2004). Semiparametric estimation of a panel data proportional hazard model with fixed effects, *Journal of Econometrics*, 119: 155–198.

Hsiao, C. (2003). *Analysis of Panel Data* (2nd. edition), Combridge: Cambridge University Press.

Hu, L. (2002). Estimating a censored dynamic panel data model with an application to earnings dynamics, *Econometrica*, 70: 2499–2517.

Kniesner, T. and Li, Q. (2002). Semiparametric panel data models with heterogeneous dynamic adjustment: Theoretical consideration and an application to labor supply, *Empirical Economics*, 27: 131–148.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model, *Econometrica*, 65: 1335–1364.

Lee, M.J. (1999). A root-N consistent semiparametric estimator for fixed effects binary response panel data, *Econometrica*, 67: 427–433.

Lee, M.J. (2001), First-Difference estimator for panel censored-selection models, *Economics Letters*, 70: 43–49.

Li, Q. and Hsiao, C. (1998). Testing serial correlation in semiparametric panel data models, *Journal of Econometrics*, 87: 207–237.

Li, Q. and Wang, S. (1998). A simple bootstrap test for a parametric regression functional form, *Journal of Econometrics*, 87: 145–165.

Linton, O. and Nielsen, J.P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration, *Biometrika*, 83: 93–100.

Manski, C. (1987). Semiparametric analysis of random effects linear models from binary panel data, *Econometrica*, 55: 357–362.

Newey, W.K. (1994). Kernel estimation of partial means in a general variance estimator, *Econometric Theory*, 10: 233–253.

Newey, W.K. (1997). Convergence rates and asymptotic normality for series estimators, *Journal of Econometrics*, 79: 147–168.

Robinson, P. (1988). Root-N-consistent semiparametric regression, *Econometrica*, 56: 931–954.

Shen, X. (1997). On methods of sieves and penalization, *Annals of Statistics*, 25: 2555–2591.

Tjostheim, D. and Auestad, B.H. (1994). Nonparametric identification of nonlinear time series: projections, *Journal of American Statistical Association*, 89:1398–1409.

Van den Berg, G. (2001). Duration models: specification, identification, and multiple durations, in *Handbook of Econometrics*, Volume 5, Amsterdam: North-Holland, 3381–3460.

Wooldridge, J.M. (1995). Selection corrections for panel data models under conditional mean independence assumptions, *Journal of Econometrics*, 68: 115–132.

Chapter 15
Panel Data Modeling and Inference: A Bayesian Primer

Siddhartha Chib

15.1 Introduction

In this chapter we discuss how Bayesian methods are used to model and analyze panel data. As in other areas of econometrics and statistics, the growth of Bayesian ideas in the panel data setting has been aided by the revolutionary developments in *Markov chain Monte Carlo* (MCMC) methods. These methods, applied creatively, allow for the sophisticated modeling of continuous, binary, censored, count and multinomial responses under weak assumptions. The purpose of this largely self-contained chapter is to summarize the various modeling possibilities and to provide the associated inferential techniques for conducting the prior-posterior analyses.

The apparatus we outline in this chapter relies on some powerful and easily implementable Bayesian precepts (for a textbook discussion of Bayesian methods, see Congdon (2001)). One theme around which much of the discussion is organized is *hierarchical prior modeling* (Lindley and Smith (1972)) which allows the researcher to model cluster-specifc heterogeneity (and its dependence on cluster-specific covariates) through random effects and random-coefficients in various interesting ways. Another theme is the use of the general approaches of Albert and Chib (1993) and Chib (1992) for dealing with binary, ordinal and censored outcomes. A third theme is the use of flexible and robust families of parametric distributions to represent sampling densities and prior distributions. A fourth theme is the comparison of alternative clustered data models via marginal likelihoods and Bayes factors, calculated via the method of Chib (1995). A final theme is the use of MCMC methods (Gelfand and Smith (1990), Tierney (1994), Chib and Greenberg (1995, 1996)) to sample the posterior distribution, to calculate the predictive density and the posterior distribution of the residuals, and to estimate the marginal likelihood.

Siddhartha Chib
Olin Business School, Campus Box 1133, Washington University in St. Louis, 1 Brookings Dr, St. Louis, MO 63130, e-mail: chib@wustl.edu

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

Because implementation of the Bayesian paradigm is inextricably tied to MCMC methods, we include a brief overview of MCMC methods and of certain basic results that prove useful in the derivation of the conditional densities that form the basis for model fitting by MCMC simulation methods. Methods for producing random variates from a few common distributions are also included. After these preliminaries, the chapter turns to the analysis of panel data models for continuous outcomes followed by a discussion of models and methods for binary, censored, count and multinomial outcomes. The last half of the chapter deals with the problems of an endogenous covariate, informative missingness, prediction, residual analysis and model comparison.

15.1.1 Hierarchical Prior Modeling

The Bayesian approach to panel data modeling relies extensively on the idea of a hierarchical prior which is used to model the heterogeneity in subject-specific coefficients and the distribution of the errors and the random effects. Suppose that for the ith cluster (subject) in the sample we are interested in modeling the distribution of $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$ on a continuous response y. Also suppose that $\mathbf{W}_i = (\mathbf{w}_{i1}, \ldots, \mathbf{w}_{in_i})'$ is a $n_i \times q$ matrix of observations on q covariates $\mathbf{w}_{it}$ whose effect on y is assumed to be cluster-specific. In particular, suppose that for the ith subject at the tth time point one writes

$$y_{it} = \mathbf{w}_{it}'\beta_i + \varepsilon_{it}\ ,\ i = 1, 2, \ldots, N;\ t = 1, 2, \ldots, n_i \tag{15.1}$$

or equivalently for all observations in the ith cluster

$$\begin{aligned} &\mathbf{y}_i = \mathbf{W}_i\beta_i + \varepsilon_i,\ i = 1, 2, \ldots, N \\ &\varepsilon_i \sim P \end{aligned}$$

where β_i is the cluster-specific coefficient vector and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})$ is the error distributed *marginally* with mean zero according to the distribution P (to be modeled below).

In the context of observational data, one is concerned about the presence of unobserved confounders (variables that simultaneously affect the covariates $\mathbf{w}_{it}$ and the error ε_{it}). Under such endogeneity of the covariates, $E(\varepsilon_i|\mathbf{W}_i, \beta_i)$ is not zero and the cluster-specific effects are not identified without additional assumptions and the availability of instruments. To make progress, and to avoid the latter situation, it is common to assume that the covariates $\mathbf{w}_{it}$ are strictly exogenous in the sense that ε_i is uncorrelated with $\mathbf{W}_i$ and β_i, which implies that ε_{it} is uncorrelated with past, current and future values of $\mathbf{w}_{it}$, given β_i, or in other words, that the distribution of ε_i given $(\mathbf{W}_i, \beta_i)$ is P. In the Bayesian context, this strict exogeneity assumption is not required and analysis can proceed under the weaker *sequential exogeneity* assumption wherein ε_{it} is uncorrelated with $\mathbf{w}_{it}$ given past values of $\mathbf{w}_{it}$ and β_i. Most of our analysis, in fact, is conducted under this assumption, although we do not make it explicit in the notation. There are situations, of course, where even the assumption

of sequential exogeneity is not tenable. We consider one such important case below where a time-varying binary covariate (a non-randomly assigned "treatment") is correlated with the error. We show how the Bayesian analysis is conducted when an instrument is available to model the marginal distribution of the treatment.

In practice, even when the assumption of sequential exogeneity of the covariates $\mathbf{w}_{it}$ holds, it is quite possible that there exist covariates $\mathbf{a}_i : r \times 1$ (with an intercept included) that are correlated with the random-coefficients β_i. These subject-specific covariates may be measurements on the subject at baseline (time $t = 0$) or other time-invariant covariates. In the Bayesian hierarchical approach this dependence on subject-specific covariates is modeled by a hierarchical prior. One quite general way to proceed is to assume that

$$\underbrace{\begin{pmatrix} \beta_{i1} \\ \beta_{i2} \\ \vdots \\ \beta_{iq} \end{pmatrix}}_{\beta_i} = \underbrace{\begin{pmatrix} \mathbf{a}_i' & \mathbf{0}' & \cdots & \cdots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{a}_i' & \cdots & \cdots & \mathbf{0}' \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \cdots & \mathbf{a}_i' \end{pmatrix}}_{\mathbf{A}_i} \underbrace{\begin{pmatrix} \beta_{11} \\ \beta_{22} \\ \vdots \\ \vdots \\ \beta_{qq} \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{iq} \end{pmatrix}}_{\mathbf{b}_i}$$

or in vector-matrix form

$$\beta_i = \mathbf{A}_i \beta + \mathbf{b}_i$$

where $\mathbf{A}_i$ is a $q \times k$ matrix given as $\mathbf{I}_q \otimes \mathbf{a}_i'$, $k = r \times q$, $\beta = (\beta_{11}, \beta_{22}, \ldots, \beta_{qq})$ is a $k \times 1$ dimensional vector, and $\mathbf{b}_i$ is the mean zero random effects vector (uncorrelated with $\mathbf{A}_i$ and ε_i) that is distributed according to the distribution Q. This is the second-stage of the model. It may be noted that the matrix $\mathbf{A}_i$ can be the identity matrix of order q or the zero matrix of order q. Thus, the effect of $\mathbf{a}_i$ on β_{i1} (the intercept) is measured by β_{11}, that on β_{i2} is measured by β_{22} and that on β_{iq} by β_{qq}.

In the same way, the hierarchical approach can be used to model the distributions P and Q. One way is to assume that each of these distributions belong to the (hierarchical) scale mixture of normals family. Formally, to model the distribution of ε_i, we could, for example, let

$$\varepsilon_i | \sigma^2, \lambda_i, \mathbf{\Omega}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2 \lambda_i^{-1} \mathbf{\Omega}_i)$$
$$\lambda_i \sim G$$

where $\mathbf{\Omega}_i$ is a positive-definite matrix depending perhaps on a set of unknown parameters ϕ, σ^2 is an unknown positive scale parameter, and λ_i is the random scale parameter that is drawn independently across clusters from some distribution G (say with known parameters). If for example, we assume that

$$G = \mathcal{G}\left(\frac{\nu_G}{2}, \frac{\nu_G}{2}\right)$$

where $\mathcal{G}$ denotes the gamma distribution, then the distribution of ε_i marginalized over λ_i is multivariate-t with density proportional to

$$|\Omega|^{1/2}\left(1+\frac{1}{\nu\sigma^2}\varepsilon'\Omega_i^{-1}\varepsilon\right)^{-(\nu+n_i)/2}.$$

Similarly, to model the random effects vector $\mathbf{b}_i$ we could let

$$\begin{aligned}\mathbf{b}_i|\eta_i,\mathbf{D} &\sim \mathcal{N}_q\left(\mathbf{0},\eta_i^{-1}\mathbf{D}\right)\\ \eta_i &\sim F\end{aligned}$$

where $\mathbf{D}$ is a full matrix and η_i is a positive random variable drawn independently across clusters from a distribution F.

The Bayesian hierarchical model is completed through the specification of prior densities on all the non-cluster-specific coefficients. In general terms, we let

$$(\beta,\mathbf{D},\sigma^2,\phi)\sim\pi$$

where π is some suitable parametric distribution. Interestingly, it is possible to model the prior distribution in stages by putting a prior on the parameters (hyperparameters) of π. Note that the latter distribution is a prior distribution on parameters from the different stages of the hierarchical model.

As another example of a hierarchical model, suppose that $\mathbf{X}_{1i}$ is an additional $n_i\times k_1$ matrix of observations on k_1 covariates whose effect on y is assumed to be non-cluster-specific. Now suppose that the model generating $\mathbf{y}_i$ is taken to

$$\mathbf{y}_i=\mathbf{X}_{1i}\beta_1+\mathbf{W}_i\beta_{2i}+\varepsilon_i,\ i=1,2,\ldots,N \tag{15.2}$$

where, as above, the distribution of the subject-specific β_{2i} is modeled as

$$\beta_{2i}=\mathbf{A}_i\beta_2+\mathbf{b}_i$$

with the remaining components of the model unchanged. In this hierarchical model, if $\mathbf{A}_i$ is not the zero matrix then identifiability requires that the matrices $\mathbf{X}_{1i}$ and $\mathbf{W}_i$ have no covariates in common. For example, if the first column of $\mathbf{W}_i$ is a vector of ones, then $\mathbf{X}_{1i}$ cannot include an intercept. If $\mathbf{A}_i$ is the zero matrix, however, $\mathbf{W}_i$ is typically a subset of $\mathbf{X}_{1i}$.

These two types of hierarchical Bayesian models play a large role in the Bayesian analysis of clustered data. Notice that both models share the same form. This is seen by inserting the model of the cluster-specific random coefficients into the first stage which yields

$$\begin{aligned}\mathbf{y}_i=\mathbf{X}_i\beta+\mathbf{W}_i\mathbf{b}_i+\varepsilon_i\ ,\ \varepsilon_i|\sigma^2,\lambda_i,\ \mathbf{\Omega}_i &\sim \mathcal{N}_{n_i}(\mathbf{0},\sigma^2\lambda_i^{-1}\mathbf{\Omega}_i)\\ \mathbf{b}_i|\eta_i,\mathbf{D} &\sim \mathcal{N}_q\left(\mathbf{0},\eta_i^{-1}\mathbf{D}\right)\\ \lambda_i\sim G\ ,\ \eta_i &\sim F\\ (\beta,\mathbf{D},\sigma^2) &\sim \pi\end{aligned}$$

where in the first type of hierarchical model

$$\mathbf{X}_i = \mathbf{W}_i \mathbf{A}_i$$

and in the second type of hierarchical model

$$\mathbf{X}_i = (\mathbf{X}_{1i} \;\; \mathbf{W}_i \mathbf{A}_i) \text{ with } \beta = (\beta_1 \; \beta_2) \,,$$

as is readily checked. The latter model is therefore the canonical Bayesian hierarchical model for continuous clustered data.

15.1.2 Elements of Markov Chain Monte Carlo

The basic idea behind MCMC methods is quite simple. Suppose that $\pi(\boldsymbol{\psi}|\mathbf{y}) \propto \pi(\boldsymbol{\psi}) p(\mathbf{y}|\boldsymbol{\psi})$ is the posterior density for a set of parameters $\psi \in \Re^d$ in a particular Bayesian model defined by the prior density $\pi(\boldsymbol{\psi})$ and sampling density or likelihood function $p(\mathbf{y}|\boldsymbol{\psi})$ and that interest centers on the posterior mean $\eta = \int_{\Re^d} \psi \pi(\psi|\mathbf{y}) \mathrm{d}\psi$. Now suppose that this integral cannot be computed analytically and that the dimension of the integration exceeds three or four (which essentially rules out the use of standard quadrature-based methods). In such cases one calculates the integral by Monte Carlo sampling methods. The general idea is to abandon the immediate task at hand (which is the computation of the above integral) and to ask how the posterior density $\pi(\boldsymbol{\psi}|y)$ may be sampled. The reason for changing our focus is that if we were to have the draws

$$\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(M)} \sim \pi(\boldsymbol{\psi}|\mathbf{y}) \,,$$

from the posterior density, then provided the sample is large enough, we estimate not just the above integral but also other features of the posterior density by taking those draws and forming the relevant sample-based estimates. For example, the sample average of the sampled draws is our simulation-based estimate of the posterior mean, while the quantiles of the sampled output are estimates of the posterior quantiles, with other summaries obtained in a similar manner. Under suitable laws of large numbers these estimates converge to the posterior quantities as the simulation-size becomes large. In short, the problem of computing an intractable integral is reduced to the problem of sampling the posterior density.

The sampling of the posterior distribution is, therefore, the central focus of Bayesian computation. One important breakthrough in the use of simulation methods was the realization that the sampled draws need not be independent, that simulation-consistency can be achieved with correlated draws. The fact that the sampled variates can be correlated is of immense practical and theoretical importance and is the defining characteristic of Markov chain Monte Carlo methods, popularly referred to by the acronym MCMC, where the sampled draws form a Markov chain. The idea behind these methods is simple and extremely general. In order to sample a given probability distribution, referred to as the target distribution, a suitable Markov chain is constructed with the property that its limiting, invariant distribution,

is the target distribution. Once the Markov chain has been constructed, a sample of draws from the target distribution is obtained by simulating the Markov chain a large number of times and recording its values. Within the Bayesian framework, where both parameters and data are treated as random variables and inferences about the parameters are conducted conditioned on the data, the posterior distribution of the parameters provides a natural target for MCMC methods.

Markov chain sampling methods originate with the work of Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) in statistical physics. A vital extension of the method was made by Hastings (1970) leading to a method that is now called the Metropolis–Hastings algorithm (see Chib and Greenberg (1995) for a detailed summary). This algorithm was first applied to problems in spatial statistics and image analysis (Besag (1974)). A resurgence of interest in MCMC methods started with the papers of Geman and Geman (1984) who developed an algorithm, a special case of the Metropolis method that later came to be called the Gibbs sampler, to sample a discrete distribution, Tanner and Wong (1987) who proposed a MCMC scheme involving data augmentation to sample posterior distributions in missing data problems, and Gelfand and Smith (1990) where the value of the Gibbs sampler was demonstrated for general Bayesian problems with continuous parameter spaces.

The Gibbs sampling algorithm is one of the simplest Markov chain Monte Carlo algorithms and is easy to describe. Suppose that for some grouping of the parameters into sub-blocks, say $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ (the extension to more than two blocks is straightforward), the set of full conditional densities

$$\pi_1(\boldsymbol{\psi}_1|\mathbf{y},\boldsymbol{\psi}_2) \propto p(\mathbf{y}|\boldsymbol{\psi}_1,\boldsymbol{\psi}_2)\pi(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2) \tag{15.3}$$

$$\pi_2(\boldsymbol{\psi}_2|\mathbf{y},\boldsymbol{\psi}_1) \propto p(\mathbf{y}|\boldsymbol{\psi}_1,\boldsymbol{\psi}_2)\pi(\boldsymbol{\psi}_1,\boldsymbol{\psi}_2) \tag{15.4}$$

are tractable (that is, of known form and readily sampled). Then, one cycle of the Gibbs sampling algorithm is completed by sampling each of the full conditional densities, using the most current values of the conditioning block. The Gibbs sampler in which each block is revised in fixed order is defined as follows.

Algorithm: Gibbs Sampling

1. Specify an initial value $\boldsymbol{\psi}^{(0)} = (\boldsymbol{\psi}_1^{(0)}, \boldsymbol{\psi}_2^{(0)})$:
2. Repeat for $j = 1, 2, \ldots, n_0 + G$.
   Generate $\boldsymbol{\psi}_1^{(j)}$ from $\pi_1(\boldsymbol{\psi}_1|\mathbf{y}, \boldsymbol{\psi}_2^{(j-1)})$
   Generate $\boldsymbol{\psi}_2^{(j)}$ from $\pi_2(\boldsymbol{\psi}_2|\mathbf{y}, \boldsymbol{\psi}_1^{(j)})$
3. Return the values $\{\boldsymbol{\psi}^{(n_0+1)}, \boldsymbol{\psi}^{(n_0+2)}, \ldots, \boldsymbol{\psi}^{(n_0+G)}\}$.

To illustrate the manner in which the blocks are revised, consider Fig. 15.1 which traces out a possible trajectory of the sampling algorithm under the assumption that each block consists of a single component. The contours in the plot represent the joint distribution of $\boldsymbol{\psi}$ and the labels "(0)", "(1)" etc., denote the simulated values. Note that one iteration of the algorithm is completed after both components are

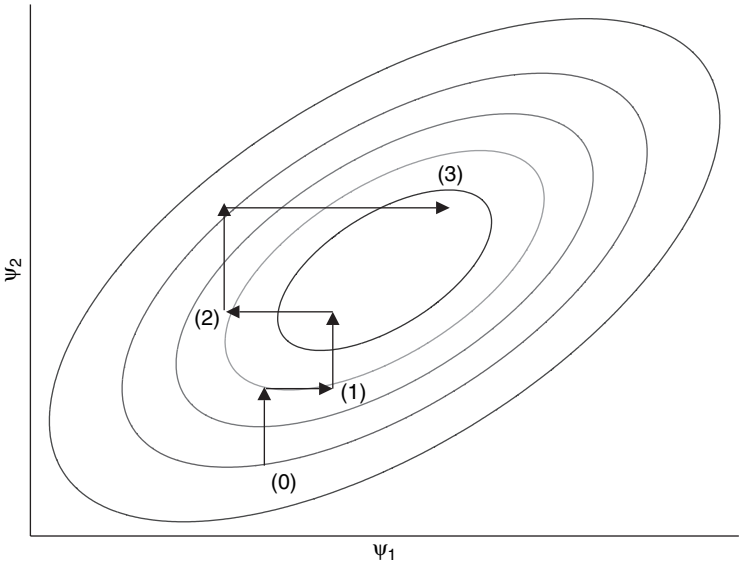

Fig. 15.1 Gibbs algorithm: An illustrative sequence of three draws

revised. Also notice that each component is revised along the direction of the coordinate axes. This feature is a source of problems if the two components are highly correlated because then the contours become compressed and movements along the coordinate axes tend to produce only small moves.

In some problems it turns out that the full conditional density cannot be sampled directly. In such cases, the intractable full conditional density is sampled via the Metropolis–Hastings (M–H) algorithm. For specificity, suppose that the full conditional density $\pi(\boldsymbol{\psi}_1|\mathbf{y}, \boldsymbol{\psi}_2)$ is intractable. Let

$$q_1(\boldsymbol{\psi}_1, \boldsymbol{\psi}_1'|\mathbf{y},\ \boldsymbol{\psi}_2)$$

denote a suitably chosen proposal density of making a transition from $\boldsymbol{\psi}_1$ to $\boldsymbol{\psi}_1'$, given the data and the values of the remaining blocks (see for example Chib and Greenberg (1995)). Then, in the first step of the jth iteration of the MCMC algorithm, given the values $\boldsymbol{\psi}_2^{(j-1)}$ of the remaining block, the updated iterate of $\boldsymbol{\psi}_1$ is drawn as follows.

Algorithm: Metropolis–Hastings for sampling an intractable $\pi_1(\boldsymbol{\psi}_1|\mathbf{y}, \boldsymbol{\psi}_2)$

1. Propose a value for $\boldsymbol{\psi}_1$ by drawing:

$$\boldsymbol{\psi}_1' \sim q_1(\boldsymbol{\psi}_1^{(j-1)}, \cdot|\mathbf{y},\ \boldsymbol{\psi}_2^{(j-1)})$$

2. Calculate the probability of move $\alpha(\boldsymbol{\psi}_1^{(j-1)}, \boldsymbol{\psi}_1' | \mathbf{y}, \boldsymbol{\psi}_2^{(j-1)})$ given by

$$\min\left\{1, \frac{\pi(\boldsymbol{\psi}_1'|\mathbf{y}, \boldsymbol{\psi}_2^{(j-1)}) q_1(\boldsymbol{\psi}_1', \boldsymbol{\psi}_1^{(j-1)}|\mathbf{y}, \boldsymbol{\psi}_2^{(j-1)})}{\pi(\boldsymbol{\psi}_1|\mathbf{y}, \boldsymbol{\psi}_2^{(j-1)}) q_1(\boldsymbol{\psi}_1^{(j-1)}, \boldsymbol{\psi}_1'|\mathbf{y}, \boldsymbol{\psi}_2^{(j-1)})}\right\}.$$

3. Set

$$\boldsymbol{\psi}_1^{(j)} = \begin{cases} \boldsymbol{\psi}_1' & \text{with prob } \alpha(\boldsymbol{\psi}_1^{(j-1)}, \boldsymbol{\psi}_1' | \mathbf{y}, \boldsymbol{\psi}_2^{(j-1)}) \\ \boldsymbol{\psi}_1^{(j-1)} & \text{with prob } 1 - \alpha(\boldsymbol{\psi}_1^{(j-1)}, \boldsymbol{\psi}_1' | \mathbf{y}, \boldsymbol{\psi}_2^{(j-1)}) \end{cases}.$$

A similar approach is used to sample $\boldsymbol{\psi}_2$ if the full conditional density of $\boldsymbol{\psi}_2$ is intractable. These algorithms are extended to more than two blocks in a straightforward manner (Chib (2001)).

15.1.3 Some Basic Bayesian Updates

We now summarize four results that appear in the development of the MCMC algorithms for the various models that are discussed below. These results provide, for the stated models, the posterior distribution of a set of parameters, conditional on the other parameters of the model. The results are stated in some generality and are specialized, as needed, in the subsequent discussion.

Result 15.1 Suppose that

$$\begin{aligned} \mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{W}_i\mathbf{b}_i + \varepsilon_i \ , \ \varepsilon_i|\sigma^2, \lambda_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i) \ , \ i \le N \\ \mathbf{b}_i|\eta_i, \mathbf{D} \sim \mathcal{N}_q(\mathbf{0}, \eta_i^{-1}\mathbf{D}) \\ \beta \sim \mathcal{N}_k(\beta_0, \mathbf{B}_0) \end{aligned}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$ is a vector of n_i observations on the dependent variable for subject i. Then marginalized over $\{\mathbf{b}_i\}$

$$\beta|\mathbf{y}, \sigma^2, \{\lambda_i\}, \{\eta_i\}, \boldsymbol{\Omega}_i, \mathbf{D} \sim \mathcal{N}_k\left\{\hat{\beta}, \mathbf{B}\right\} \tag{15.5}$$

where

$$\hat{\beta} = \mathbf{B}\left(\mathbf{B}_0^{-1}\beta_0 + \sum_{i=1}^N \mathbf{X}_i\mathbf{V}_i^{-1}\mathbf{y}_i\right), \tag{15.6}$$

$$\mathbf{B} = \left(\mathbf{B}_0^{-1} + \sum_{i=1}^N \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1} \tag{15.7}$$

and

$$\mathbf{V}_i = \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i + \eta_i^{-1}\mathbf{W}_i\mathbf{D}\mathbf{W}_i' \tag{15.8}$$

Result 15.2 Suppose that

$$\begin{aligned} \mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{W}_i\mathbf{b}_i + \varepsilon_i \ , \ \varepsilon_i|\sigma^2, \lambda_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i) \\ \mathbf{b}_i|\eta_i, \mathbf{D} \sim \mathcal{N}_q(\mathbf{0}, \eta_i^{-1}\mathbf{D}) \end{aligned}$$

Then

$$\mathbf{b}_i|\mathbf{y}_i,\beta,\sigma^2,\lambda_i,\eta_i,\mathbf{D}\sim\mathcal{N}_q\left(\hat{\mathbf{b}}_i,\mathbf{D}_i\right) \tag{15.9}$$

where

$$\hat{\mathbf{b}}_i=\sigma^{-2}\lambda_i\mathbf{D}_i\mathbf{W}_i'\boldsymbol{\Omega}_i^{-1}(\mathbf{y}_i-\mathbf{X}_i\beta) \tag{15.10}$$

and

$$\mathbf{D}_i=(\eta_i\mathbf{D}^{-1}+\sigma^{-2}\lambda_i\mathbf{W}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{W}_i)^{-1}. \tag{15.11}$$

Result 15.3 Suppose that

$$\begin{aligned}\mathbf{y}_i=\mathbf{X}_i\beta+\mathbf{W}_i\mathbf{b}_i+\varepsilon_i\ ,\ \varepsilon_i|\sigma^2,\ \lambda_i\sim\mathcal{N}_{n_i}(\mathbf{0},\ \sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i)\\ \mathbf{b}_i|\eta_i,\mathbf{D}\sim\mathcal{N}_q(\mathbf{0},\eta_i\mathbf{D})\,,\ i\le N\\ \mathbf{D}^{-1}\sim\mathcal{W}_q(\rho_0,\mathbf{R}_0)\end{aligned}$$

where $\mathcal{W}_T(\rho,\mathbf{R})$ is the Wishart distribution with density

$$c\frac{|\mathbf{W}|^{(\nu-T-1)/2}}{|\mathbf{R}|^{\nu/2}}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\mathbf{R}^{-1}\mathbf{W}\right)\right\},\quad|\mathbf{W}|>0,$$

$$c=\left(2^{\rho T/2}\pi^{T(T-1)/4}\prod_{i=1}^{T}\Gamma\left(\frac{\rho+1-i}{2}\right)\right)^{-1}$$

is the normalizing constant and $\mathbf{R}$ is a hyperparameter matrix (Roberts (2001)). Then

$$\mathbf{D}^{-1}|\{\mathbf{b}_i\},\mathbf{y},\boldsymbol{\Omega}_i,\{\lambda_i\},\{\eta_i\}=\mathbf{D}^{-1}|\{\mathbf{b}_i\},\eta_i\sim\mathcal{W}_q\left(\rho_0+N,\mathbf{R}\right) \tag{15.12}$$

where

$$\mathbf{R}=\left(\mathbf{R}_0^{-1}+\sum_{i=1}^{N}\eta_i\mathbf{b}_i\mathbf{b}_i'\right)^{-1}. \tag{15.13}$$

Result 15.4 Suppose that

$$\begin{aligned}\mathbf{y}_i=\mathbf{X}_i\beta+\mathbf{W}_i\mathbf{b}_i+\varepsilon_i\ ,\ \varepsilon_i|\sigma^2,\lambda_i\sim\mathcal{N}_{n_i}(\mathbf{0},\sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i)\,,\ i\le N\\ \sigma^2\sim\mathcal{IG}\left(\frac{\nu_0}{2},\frac{\delta_0}{2}\right)\end{aligned}$$

where $\mathcal{IG}(a,b)$ is the inverse-gamma distribution with density $\pi(\sigma^2|a,b)\propto\left(\sigma^2\right)^{-a+1}\exp\left(-b/\sigma^2\right)$. Then

$$\sigma^2|\mathbf{y},\beta,\{\mathbf{b}_i\},\{\lambda_i\}\sim\mathcal{IG}\left(\frac{\nu_0+\sum n_i}{2},\frac{\delta_0+\delta}{2}\right) \tag{15.14}$$

where

$$\delta=\sum_{i=1}^{N}\lambda_i\mathbf{e}_i{}'\boldsymbol{\Omega}_i^{-1}\mathbf{e}_i \tag{15.15}$$

and

$$\mathbf{e}_i = (\mathbf{y}_i - \mathbf{X}_i\beta - \mathbf{W}_i\mathbf{b}_i)$$

15.1.4 Basic Variate Generators

In the application of MCMC methods it often occurs that the simulation of the given target distribution is reduced to a sequence of simulations from standard and familiar univariate and multivariate distributions. With that in mind, we present simulation routines for the distributions that are encountered in the sequel.

Gamma Variate: To obtain ψ from $\mathcal{G}(\alpha,\beta)$ with density proportional to $\psi^{\alpha-1}\exp(-\beta\psi)$, we draw θ from $\mathcal{G}(\alpha,1)$ and set $\psi = \theta/\beta$. A draw of a chi-squared variate χ^2_ν with ν degrees of freedom is obtained by drawing from a $\mathcal{G}(\alpha/2,1/2)$ distribution.

Inverse-Gamma Variate: A random variable that follows the inverse-gamma distribution $\mathcal{IG}(\alpha,\beta)$ is equal in distribution to the inverse of random variable that follows the $\mathcal{G}(\alpha,\beta)$ distribution. Therefore, an inverse-gamma variate is obtained by drawing θ from $\mathcal{G}(\alpha,\beta)$ and setting $\psi = 1/\theta$.

Truncated Normal Variate: A variate from

$$\psi \sim \mathcal{TN}_{(a,b)}(\mu,\sigma^2),$$

a univariate normal distribution truncated to the interval (a,b), is obtained by the inverse-cdf method. The distribution function of the truncated normal random variable is

$$F(t) = \begin{cases} 0 & \text{if } \psi < a \\ \frac{1}{p_2-p_1}\left(\Phi(\frac{t-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})\right) & \text{if } a < \psi < b \\ 1 & \text{if } b < \psi \end{cases} \tag{15.16}$$

where

$$p_1 = \Phi\left(\frac{a-\mu}{\sigma}\right)\ ;\ p_2 = \Phi\left(\frac{b-\mu}{\sigma}\right)$$

Therefore, if U is uniform on $(0,1)$, then

$$\psi = \mu + \sigma\Phi^{-1}(p_1 + U(p_2 - p_1)) \tag{15.17}$$

is the required draw. Here Φ^{-1} is the inverse-cdf of the standard normal distribution and can be computed by the method of Page (1977).

Multivariate Normal Vector: To obtain a random vector $\boldsymbol{\psi}$ from $\mathcal{N}_k(\mu,\boldsymbol{\Omega})$, we draw θ from $\mathcal{N}_k(\mathbf{0},\mathbf{I}_k)$ and set $\boldsymbol{\psi} = \mu + \mathbf{L}\boldsymbol{\theta}$ where $\boldsymbol{\Omega} = \mathbf{L}\mathbf{L}'$.

Wishart Matrix: To obtain a random positive-definite matrix W from $\mathcal{W}_T(v,\mathbf{R})$, one first generates the random lower triangular matrix $\mathbf{T} = (t_{ij})$, such that

$$t_{ii} \sim \sqrt{\chi^2_{\nu-i+1}} \text{ and } t_{ij} \sim \mathcal{N}(0,1)$$

Then the quantity

$$\mathbf{W} = \mathbf{LTT}'\mathbf{L}'$$

where $\mathbf{R} = \mathbf{LL}'$ is the required draw.

15.2 Continuous Responses

As discussed in Sect. 1.1, Bayesian hierarchical modeling of subject-specific coefficients leads to the canonical model for unbalanced continuous outcomes

$$\begin{aligned}
\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{W}_i\mathbf{b}_i + \varepsilon_i \ , \ \varepsilon_i|\sigma^2, \lambda_i, \mathbf{\Omega}_i &\sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\lambda_i^{-1}\mathbf{\Omega}_i) \\
\mathbf{b}_i|\eta_i, \mathbf{D} &\sim \mathcal{N}_q\left(\mathbf{0}, \eta_i^{-1}\mathbf{D}\right) \\
\lambda_i \sim G \ , \ \eta_i &\sim F \\
(\beta, \mathbf{D}, \sigma^2) &\sim \pi
\end{aligned}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$ is the data on the ith individual over the n_i time periods, $\mathbf{W}_i$ is a set of variables whose effect $\mathbf{b}_i$ is assumed to heterogenous, $\mathbf{X}_i$ is a set of raw covariates or the matrix $\mathbf{W}_i\mathbf{A}_i$ or $(\mathbf{X}_{1i} \ \mathbf{W}_i\mathbf{A}_i)$ if the model is derived from a hierarchical specification in which the heterogeneity depends on cluster-specific covariates $\mathbf{A}_i$.

There are many ways to proceed from this point. If G and F are degenerate at one, we get the Gaussian-Gaussian model. If we assume that

$$G = \mathcal{G}\left(\frac{\nu_G}{2}, \frac{\nu_G}{2}\right)$$

and

$$F = \mathcal{G}\left(\frac{\nu_F}{2}, \frac{\nu_F}{2}\right)$$

then the distributions of ε_i and $\mathbf{b}_i$ marginalized over λ_i and η_i are multivariate student-t with ν_G and ν_F degrees of freedom, respectively. This model may be called the Student–Student model. Other models are obtained by making specific assumptions about the form of $\mathbf{\Omega}_i$. For example, if ε_i is assumed to to be serially correlated according to say an ARMA process, then $\mathbf{\Omega}_i$ is the covariance matrix of the assumed ARMA process. The distribution π is typically specified in the same way, regardless of the distributions adopted in other stages of the model. Specifically, it is common to assume that the parameters $(\beta, \mathbf{D}, \sigma^2)$ are apriori mutually independent with

$$\beta \sim \mathcal{N}_k(\beta_0, B_0); \sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right); \mathbf{D}^{-1} \sim \mathcal{W}_p(\rho_0, \mathbf{R}_0)$$

15.2.1 Gaussian–Gaussian Model

To see how the analysis may proceed, consider the model in which the distributions of the error and the random-effects are both Gaussian. In particular, suppose that

$$\varepsilon_i|\sigma^2 \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\mathbf{\Omega}_i),$$

$$\mathbf{b}_i|\mathbf{D} \sim \mathcal{N}_q(\mathbf{0}, \mathbf{D}) \ , i \leq N$$

where the matrix $\mathbf{\Omega}_i$ is assumed to be known. Under these assumptions the joint posterior of all the unknowns, including the random effects $\{\mathbf{b}_i\}$, is given by

$$\pi(\beta, \{\mathbf{b}_i\}, \mathbf{D}^{-1}, \sigma^2|\mathbf{y}) = \pi(\beta, \{\mathbf{b}_i\}, \mathbf{D}^{-1}, \sigma^2) \prod_{i=1}^{N} f(\mathbf{y}_i|\beta, \mathbf{b}_i, \sigma^2) p(\mathbf{b}_i|\mathbf{D}). \quad (15.18)$$

Wakefield, Smith, Racine Poon and Gelfand (1994) propose a Gibbs MCMC approach for sampling the joint posterior distribution based on full blocking (i.e., sampling each block of parameters from their full conditional distribution). This blocking scheme is not very desirable because the random effects and the fixed effects β tend to be highly correlated and treating them as separate blocks creates problems with mixing (Gelfand, Sahu and Carlin (1995)).

To deal with this problem, (Chib and Carlin (1999)) suggest a number of reduced blocking schemes. One of the simplest proceeds by noting that β and $\{\mathbf{b}_i\}$ can be sampled in one block by the method of composition: first sampling β marginalized over $\{\mathbf{b}_i\}$ and then sampling $\{\mathbf{b}_i\}$ conditioned on β. What makes reduced blocking possible is the fact that the conditional distribution of the outcomes marginalized over $\mathbf{b}_i$ is normal which can be combined with the assumed normal prior on β in the usual way. In particular,

$$f(\mathbf{y}_i|\beta, \mathbf{D}, \sigma^2) = \int f(\mathbf{y}_i|\beta, \mathbf{b}_i, \sigma^2) g(\mathbf{b}_i|\mathbf{D}) \mathrm{d}\mathbf{b}_i$$

$$\propto |\mathbf{V}_i|^{-1/2} \exp\{(-1/2)(\mathbf{y}_i - \mathbf{X}_i\beta)' \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\beta)\} \ ,$$

where $\mathbf{V}_i = \sigma^2\mathbf{\Omega}_i + \mathbf{W}_i\mathbf{D}\mathbf{W}_i'$, which, from Result 15.1, leads to the conditional posterior of β (marginalized over $\{\mathbf{b}_i\}$).

The rest of the algorithm follows the steps of Wakefield et al. (1994). In particular, the sampling of the random effects is from independent normal distributions that are derived by treating $(\mathbf{y}_i - \mathbf{X}_i\beta)$ as the "data," $\mathbf{b}_i$ as the regression coefficient and $\mathbf{b}_i \sim \mathcal{N}_q(0, \mathbf{D})$ as the prior and applying Result 15.2. Next, conditioned on $\{\mathbf{b}_i\}$, the full conditional distribution of $\mathbf{D}^{-1}$ becomes independent of $\mathbf{y}$ and is obtained by combining the Wishart prior distribution of $\mathbf{D}^{-1}$ with the normal distribution of $\{\mathbf{b}_i\}$ given $\mathbf{D}^{-1}$. The resulting distribution is Wishart with updated parameters obtained from Result 15.3. Finally, Result 15.4 yields the full-conditional distribution of σ^2. *In applying these results, λ_i and η_i are set equal to one in all the expressions.*

Algorithm: Gaussian–Gaussian Panel (Wakefield et al. (1994) and Chib and Carlin (1999))

1. Sample

(a.) $$\beta|\mathbf{y},\sigma^2,\mathbf{D}\sim\mathscr{N}_k\left(\hat{\beta},\mathbf{B}\right)$$

(b.) $$\mathbf{b}_i|\mathbf{y},\beta,\sigma^2,\mathbf{D}\sim\mathscr{N}_q\left(\hat{\mathbf{b}}_i,\mathbf{D}_i\right),i\leq N$$

2. Sample
$$\mathbf{D}^{-1}|\mathbf{y},\beta,\{\mathbf{b}_i\},\sigma^2\sim\mathscr{W}_q\{\rho_0+N,\mathbf{R}\}$$

3. Sample
$$\sigma^2|\mathbf{y},\beta,\{\mathbf{b}_i\},\mathbf{D}\sim\mathscr{IG}\left(\frac{\nu_0+\sum n_i}{2},\frac{\delta_0+\delta}{2}\right)$$

4. Goto 1

15.2.1.1 Example

As an illustration, we consider data from a clinical trial on the effectiveness of two antiretroviral drugs (didanosine or ddI and zalcitabine or ddC) in 467 persons with advanced HIV infection. The response variable y_{ij} for patient i at time j is the square root of the patient's CD4 count, a seriological measure of immune system health and prognostic factor for AIDS-related illness and mortality. The data set records patient CD4 counts at study entry and again at 2, 6, 12, and 18 months after entry, for the ddI and ddC groups, respectively.

The model is formulated as follows. If we let $\mathbf{y}_i$ denote a n_i vector of responses across time for the ith patient, then following the discussion in Carlin and Louis (2000), suppose

$$\begin{aligned}\mathbf{y}_i|\beta,\mathbf{b}_i,\sigma^2&\sim\mathscr{N}_{n_i}(\mathbf{X}_i\beta+\mathbf{W}_i\mathbf{b}_i,\sigma^2\boldsymbol{\Omega}_i),\ \boldsymbol{\Omega}_i=\mathbf{I}_{n_i}\\ \mathbf{b}_i|\mathbf{D}&\sim\mathscr{N}_2(\mathbf{0},\mathbf{D}),\ i\leq 467,\end{aligned}\tag{15.19}$$

where the jth row of the patient i's design matrix $\mathbf{W}_i$ takes the form $\mathbf{w}_{ij}=(1,t_{ij})$, t_{ij} belongs to the set $\{0,2,6,12,18\}$ and the fixed design matrix $\mathbf{X}_i$ is obtained by horizontal concatenation of $\mathbf{W}_i$, $d_i\mathbf{W}_i$ and $a_i\mathbf{W}_i$, where d_i is a binary variable indicating whether patient i received ddI ($d_i=1$) or ddC ($d_i=0$), and a_i is a binary variable indicating if the patient was diagnosed as having AIDS at baseline ($a_i=1$) or not ($a_i=0$).

The prior distribution of $\beta:6\times 1$ is assumed to be $\mathscr{N}_6(\beta_0,\mathbf{B}_0)$ with

$$\begin{aligned}\beta_0&=(10,0,0,0,-3,0),\quad\text{and}\\ \mathbf{B}_0&=\text{Diag}(2^2,1^2,(.1)^2,1^2,1^2,1^2),\end{aligned}$$

while that on $\mathbf{D}^{-1}$ is taken to be Wishart $W(\mathbf{R}_0/\rho_0,2,\rho_0)$ with $\rho_0=24$ and $\mathbf{R}_0=\text{diag}(.25,16)$. Finally, σ^2 is apriori assumed to follow the inverse-gamma distribution

$$\sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right),$$

with $\nu_0 = 6$ and $\delta_0 = 120$ (which imply a prior mean and standard deviation both equal to 30).

The MCMC simulation is run for 5000 cycles beyond a burn-in of a 100 cycles. The simulated values by iteration for each of the ten parameters are given in Fig. 15.2. Except for the parameters that are approximately the same, the sampled paths of the parameters are clearly visible and display little correlation.

These draws from the posterior distribution are used to produce different summaries of the posterior distribution. In Fig. 15.3 we report the marginal posterior distributions in the form of histogram plots. We see that three of the regression parameters are centered at zero, that D_{11} is large and D_{22} (which is the variance of the time-trend random effect) is small.

15.2.2 Robust Modeling of b_i: Student–Student and Student-Mixture Models

We now discuss models in which the error distribution of the observations in the *i*th cluster is multivariate-t and the distribution of $\mathbf{b}_i$ is modeled as multivariate-t or a mixture of normals. To begin, consider the student-student model

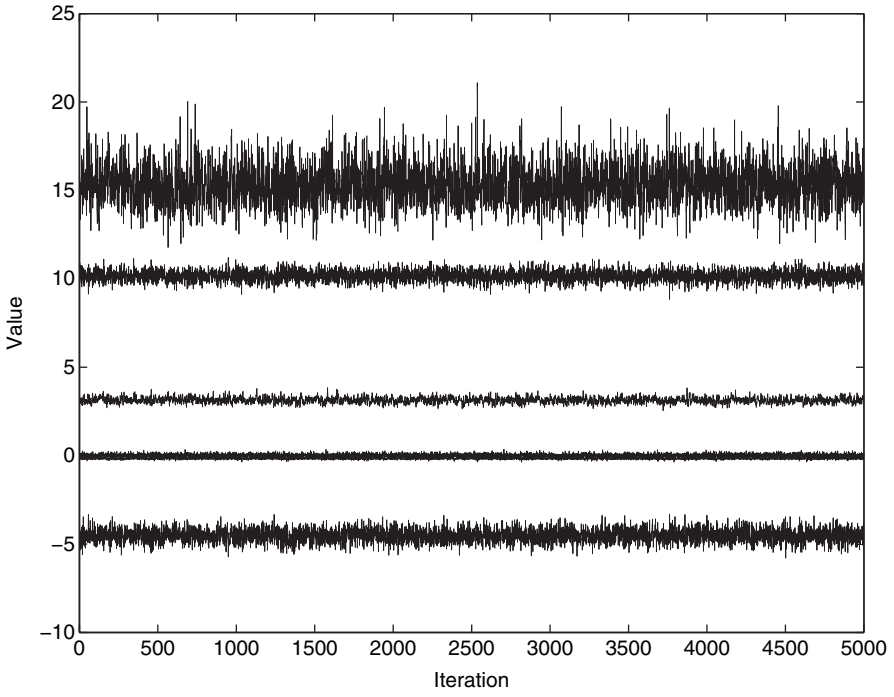

Fig. 15.2 Aids clustered data: Simulated values by iteration for each of ten parameters

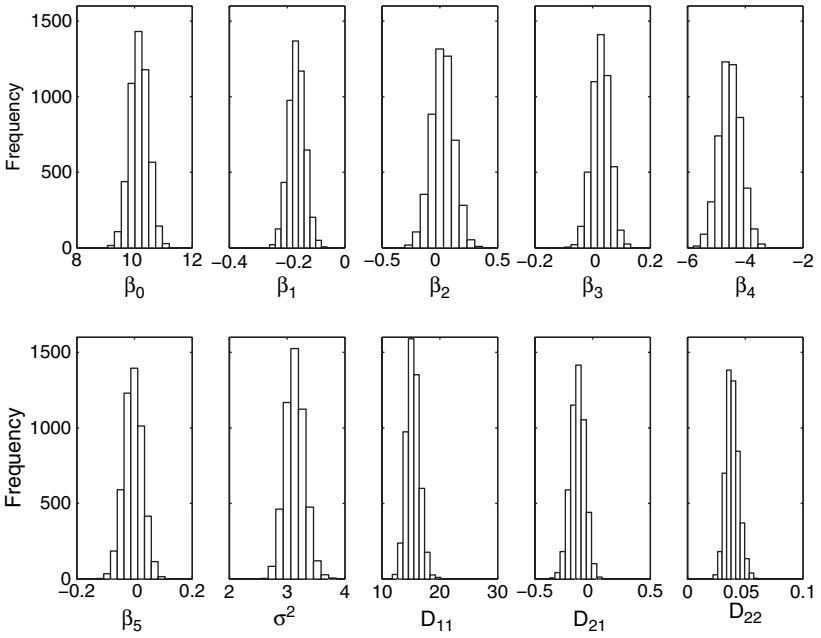

Fig. 15.3 Aids clustered data: Marginal posterior distributions of parameters based on 5000 MCMC draws

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{W}_i\mathbf{b}_i + \varepsilon_i,\ \varepsilon_i|\sigma^2, \lambda_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\lambda_i^{-1}\mathbf{\Omega}_i)$$
$$\mathbf{b}_i|\eta_i, \mathbf{D} \sim \mathcal{N}_q(\mathbf{0}, \eta_i^{-1}\mathbf{D}); i \leq N$$
$$\lambda_i \sim \mathcal{G}\left(\frac{\nu_G}{2}, \frac{\nu_G}{2}\right);\ \eta_i \sim \mathcal{G}\left(\frac{\nu_F}{2}, \frac{\nu_F}{2}\right)$$
$$\beta \sim \mathcal{N}_k(\beta_0, \mathbf{B}_0);\ \sigma^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right);\ \mathbf{D}^{-1} \sim \mathcal{W}_p(\rho_0, \mathbf{R}_0)$$

This model is easily analyzed by including λ_i and η_i, $i \leq N$, in the sampling. In that case, we follow the Gaussian–Gaussian MCMC algorithm, except that each step is implemented conditioned on $\{\lambda_i\}$ and $\{\eta_i\}$ and two new steps are added in which $\{\lambda_i\}$ and $\{\eta_i\}$ are sampled. The quantities that go into forming the various parameters in these updates are all obtained from the results of Sect. 1.3.

Algorithm: Student–Student Panel

1. Sample

(a) $$\beta|\mathbf{y}, \sigma^2, \mathbf{D}, \{\lambda_i\}, \{\eta_i\} \sim \mathcal{N}_k\left(\hat{\beta}, \mathbf{B}\right)$$

(b) $$\mathbf{b}_i|\mathbf{y}, \beta, \sigma^2, \mathbf{D}, \lambda_i, \eta_i \sim \mathcal{N}_q\left(\hat{\mathbf{b}}_i, \mathbf{D}_i\right),$$

2. Sample

(a)

$$\lambda_i|\mathbf{y},\beta,\{\mathbf{b}_i\},\sigma^2 \sim \mathscr{G}\left(\frac{\nu_G+n_i}{2},\frac{\nu_G+\sigma^{-2}\mathbf{e}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{e}_i}{2}\right),$$

(b)

$$\eta_i|\mathbf{b}_i,\mathbf{D} \sim \mathscr{G}\left(\frac{\nu_F+q}{2},\frac{\nu_F+\mathbf{b}_i'\mathbf{D}^{-1}\mathbf{b}_i}{2}\right), i \leq N$$

3. Sample

$$\mathbf{D}^{-1}|\mathbf{y},\beta,\{\mathbf{b}_i\},\sigma^2,\{\lambda_i\},\{\eta_i\} \sim \mathscr{W}_q\{\rho_0+N,\mathbf{R}\}$$

4. Sample

$$\sigma^2|\mathbf{y},\beta,\{\mathbf{b}_i\},\mathbf{D},\{\lambda_i\},\{\eta_i\} \sim \mathscr{IG}\left(\frac{\nu_0+\sum n_i}{2},\frac{\delta_0+\delta}{2}\right)$$

5. Goto 1

Another possibility is to assume that $\mathbf{b}_i$ is drawn from a finite mixture of Gaussian distributions. For example, one may assume that $\mathbf{b}_i \sim q_1\mathscr{N}(\mathbf{0},\mathbf{D}_1)+q_2\mathscr{N}(\mathbf{0},\mathbf{D}_2=\eta\mathbf{D}_1)$ where $\eta>1$ and q_j is the probability of drawing from the jth component of the mixture. Chen and Dunson (2003), for example, use a particular mixture prior in which one of the component random effects variances can be zero, which leads to a method for determing if the particular effect is random. Like any Bayesian analysis of a mixture model, analysis exploits the hierarchical representation of the mixture distribution:

$$\begin{aligned}\mathbf{b}_i|s_i=j &\sim \mathscr{N}(\mathbf{0},\mathbf{D}_j)\\ \Pr(s_i=j) &= q_j\ , j=1,2\end{aligned}$$

where $s_i=\{1,2\}$ is a latent population indicator variable. The MCMC based fitting of this Gaussian-mixture model proceeds by sampling the posterior distribution

$$\begin{aligned}\pi(\beta,\{\mathbf{b}_i\},\mathbf{D}_1^{-1},\sigma^2,\{\lambda_i\},\eta,\{s_i\},q|\mathbf{y}) &= \pi(\beta,\{\mathbf{b}_i\},\mathbf{D}_1^{-1},\sigma^2,\\ &\quad \{\lambda_i\},\eta,\{s_i\})f(\mathbf{y}|\,\beta,\{\mathbf{b}_i\},\sigma^2,\{\lambda_i\})\end{aligned}$$

$$=\pi(\beta)\pi(\mathbf{D}^{-1})\pi(\sigma^2)\pi(\lambda)\pi(q)\prod_{i=1}^{N}f(\mathbf{y}_i|\beta,\mathbf{b}_i,\sigma^2,\lambda_i)p(\mathbf{b}_i|s_i,\mathbf{D}_{s_i})p(s_i|q)p(\lambda_i)$$

where the prior on η is (say) inverse-gamma and that of $q=(q_1,q_2)$ a Dirichlet with density proportional to $q_1^{m_{10}-1}q_2^{m_{20}-1}$ where the hyper-parameters m_{10} and m_{20} are known. This posterior density is sampled with some minor modifications of the Student–Student algorithm. Steps 1 and 2 are now conditioned on $\{s_i\}$; as a result $\mathbf{V}_i$ in the updates is replaced by $\mathbf{V}_{s_i}=\sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i+\mathbf{W}_i\mathbf{D}_{s_i}\mathbf{W}_i'$ and $\mathbf{D}_i$ by $\mathbf{D}_{s_i}^*=(\mathbf{D}_{s_i}^{-1}+\sigma^{-2}\lambda_i\mathbf{W}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{W}_i)^{-1}$. Step 3 is now the sampling of $\mathbf{D}_1^{-1}$ where the sum

over the outer-product of the $\mathbf{b}_i$'s is replaced by $\sum_{i:s_i=1} \mathbf{b}_i\mathbf{b}_i' + \eta^{-1}\sum_{i:s_i=2} \mathbf{b}_i\mathbf{b}_i'$. Steps 4 and 5 are unchanged. Finally, two new steps are inserted: Step 6 for sampling η and Step 6 for sampling q. Each of these steps is straightforward. In Step 6 we sample η from an updated inverse-gamma distribution based on those $\mathbf{b}_i$ that are associated with population 2; the update is therefore from the model $\mathbf{b}_i|\eta, \mathbf{D}_1 \sim \mathcal{N}(\mathbf{0}, \eta\mathbf{D}_1)$, $\eta \sim IG(a_0/2, b_0/2)$ which leads to an inverse-gamma distribution. The updated distribution of q in Step 7 is easily seen to be Dirichlet with parameters $m_{10}+ m_1$ and $m_{20}+m_2$, respectively, where m_j are the total number of observations ascribed to population j in that iteration of the MCMC sampling.

15.2.3 Heteroskedasticity

The methods described above are readily adapted to deal with heteroskedasticity in the observation error process by parameterizing the error covariance matrix $\sigma^2\mathbf{\Omega}_i$. Instead of assuming that $\varepsilon_i|\sigma^2 \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\mathbf{\Omega}_i)$, we assume

$$\varepsilon_i|\sigma_i^2 \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma_i^2\mathbf{I}_{n_i})$$

where σ_i^2 can be modeled hierarchically by assuming that

$$\sigma_i^2|\delta_0 \sim IG\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right)$$
$$\delta_0 \sim G\left(\frac{\nu_{00}}{2}, \frac{\delta_{00}}{2}\right)$$

a specification that appears in Basu and Chib (2003). In the first stage of this prior specification, one assumes that conditioned on the scale of the inverse-gamma distribution, σ_i^2 is inverse-gamma and then the scale is in turn allowed to follow a gamma distribution. The fitting of this model is quite similar to the fitting of the Gaussian–Gaussian model except that $\sigma^2\mathbf{\Omega}_i$ is replaced by $\sigma_i^2\mathbf{I}_{n_i}$ in Steps 1 and 2, $\mathbf{\Omega}_i$ is replaced by $\mathbf{I}_{n_i}$ in Steps 3 and 4 is modified and a new Step 5 is inserted for the sampling of δ_0.

Algorithm: Gaussian–Gaussian Hetroskedastic Panel (Basu and Chib (2003))

1. Sample

   (a)
   $$\beta|\mathbf{y}, \{\sigma_i^2\}, \mathbf{D} \sim \mathcal{N}_k\left(\hat{\beta}, \mathbf{B}\right)$$

   (b)
   $$\mathbf{b}_i|\mathbf{y}, \beta, \{\sigma_i^2\}, \mathbf{D} \sim \mathcal{N}_q\left(\hat{\mathbf{b}}_i, \mathbf{D}_i\right), i \le N$$

2. Sample
   $$\mathbf{D}^{-1}|\mathbf{y}, \beta, \{b_i\}, \sigma^2 \sim \mathcal{W}_q\{\rho_0 + N, \mathbf{R}\}$$

3. Sample

$$\sigma_i^2|\mathbf{y},\beta,\{\mathbf{b}_i\},\mathbf{D},\delta_0 \sim \mathscr{IG}\left(\frac{\nu_0+n_i}{2},\frac{\delta_0+\parallel \mathbf{y}_i-\mathbf{X}_i\beta-\mathbf{W}_i\mathbf{b}_i\parallel^2}{2}\right)$$

4. Sample

$$\delta_0|\sigma_i^2 \sim \mathscr{G}\left(\frac{\nu_0+\nu_{00}}{2},\frac{\sigma_i^{-2}+\delta_{00}}{2}\right)$$

5. Goto 1

15.2.4 Serial Correlation

To deal with the possibility of serial correlation in models with multivariate-t error and random effects distributions we now assume that

$$\varepsilon_i|\lambda_i,\phi \sim \mathscr{N}_{n_i}(\mathbf{0},\sigma^2\lambda_i^{-1}\mathbf{\Omega}_i)$$

where $\mathbf{\Omega}_i=\mathbf{\Omega}_i(\phi)$ is a $n_i\times n_i$ covariance matrix that depends on a set of p parameters $\phi=(\phi_1,\ldots,\phi_p)$. Typically, one will assume that the errors follow a low-dimensional stationary ARMA process and the matrix $\mathbf{\Omega}_i$ will then be the covariance matrix of the n_i errors. In that case, ϕ represents the parameters of the assumed ARMA process. The fitting of this model by MCMC methods is quite straightforward. The one real new step is the sampling of ϕ by the M–H algorithm along the lines of Chib and Greenberg (1994).

Algorithm: Student–Student Correlated Error Panel

1. Sample

   (a)

$$\beta|\mathbf{y},\sigma^2,\mathbf{D},\{\lambda_i\},\phi \sim \mathscr{N}_k\left(\hat{\beta},\mathbf{B}\right)$$

   (b)

$$\mathbf{b}_i|\mathbf{y},\beta,\sigma^2,\mathbf{D},\{\lambda_i\},\{\eta_i\},\phi \sim \mathscr{N}_q\left(\hat{\mathbf{b}}_i,\mathbf{D}_i\right),$$

2. Sample

   (a)

$$\lambda_i|\mathbf{y},\beta,\{\mathbf{b}_i\},\sigma^2,\phi \sim \mathscr{G}\left(\frac{\nu_G+n_i}{2},\frac{\nu_G+\sigma^{-2}\mathbf{e}_i'\mathbf{\Omega}_i^{-1}\mathbf{e}_i}{2}\right),$$

   (b)

$$\eta_i|\mathbf{b}_i,\mathbf{D} \sim \mathscr{G}\left(\frac{\nu_F+q}{2},\frac{\nu_F+\mathbf{b}_i'\mathbf{D}^{-1}\mathbf{b}_i}{2}\right), i\le N$$

3. Sample

$$\mathbf{D}^{-1}|\mathbf{y},\beta,\{\mathbf{b}_i\},\sigma^2,\{\eta_i\},\phi \sim \mathscr{W}_q\{\rho_0+N,\mathbf{R}\}$$

4. Sample

$$\sigma^2|\mathbf{y},\beta,\{\mathbf{b}_i\},\mathbf{D},\phi \sim \mathscr{IG}\left(\frac{\nu_0+\sum n_i}{2},\frac{\delta_0+\delta}{2}\right)$$

5. Sample

$$\phi|\mathbf{y},\beta,\{\mathbf{b}_i\},\mathbf{D},\sigma^2,\{\lambda_i\} \propto \pi(\phi)\prod_{i=1}^{n}\mathscr{N}\left(\mathbf{y}_i|\mathbf{X}_i\beta+\mathbf{W}_i\mathbf{b}_i,\sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i\right)$$

6. Goto 1

In the sampling of ϕ in the above algorithm we use the tailored proposal density as suggested by Chib and Greenberg (1994). Let

$$\hat{\phi}=\arg\max_{\phi}\ln\underbrace{\left\{\pi(\phi)\prod_{i=1}^{n}\mathscr{N}\left(\mathbf{y}_i|\mathbf{X}_i\beta+\mathbf{W}_i\mathbf{b}_i,\sigma^2\lambda_i^{-1}\boldsymbol{\Omega}_i\right)\right\}}_{g(\phi)}$$

be the conditional mode of the full conditional of ϕ that is found by (say) a few steps of the Newton–Raphson algorithm, and let $\mathbf{V}$ be the symmetric matrix obtained by inverting the negative of the Hessian matrix (the matrix of second derivatives) of $\ln g(\phi)$ evaluated at $\hat{\phi}$. Then, our proposal density is given by

$$q(\phi|\mathbf{y},\beta,\{\mathbf{b}_i\},\mathbf{D},\sigma^2)=t_p(\phi|\hat{\phi},\mathbf{V},\nu)$$

a multivariate-t density with mean $\hat{\phi}$, dispersion matrix $\mathbf{V}$ and ν degrees of freedom. In this M–H step, given the current value ϕ, we now generate a proposal value ϕ' from this multivariate-t density and accept or reject with probability of move

$$\alpha(\phi,\phi'|\mathbf{y},\beta,\{\mathbf{b}_i\},\mathbf{D},\sigma^2)=\min\left\{1,\frac{g(\phi')}{g(\phi)}\frac{t_p(\phi|\hat{\phi},\mathbf{V},\nu)}{t_p(\phi'|\hat{\phi},\mathbf{V},\nu)}\right\}$$

If the proposal value is rejected we stay at the current value ϕ and move to Step 1 of the algorithm. As before, by setting λ_i and η_i to one we get the Gaussian–Gaussian version of the autoregressive model.

15.3 Binary Responses

Consider now the situation in which the response variable is binary $(0,1)$ and the objective is to fit a panel model with random effects. The classical analysis of such models (under the probit link) was pioneered by Chamberlain (1980), Heckman (1981) and Butler and Moffitt (1982).

Suppose that for the ith individual at time t, the probability of observing the outcome $y_{it}=1$, conditioned on the random effect $\mathbf{b}_i$, is given by

$$\Pr(y_{it}=1|\mathbf{b}_i)=\Phi(\mathbf{x}_{it}'\beta+\mathbf{w}_{it}'\mathbf{b}_i),$$

where Φ is the cdf of the standard normal distribution, and $\mathbf{b}_i|\mathbf{D}\sim\mathcal{N}_p(\mathbf{0},\mathbf{D})$ independent of $\mathbf{x}_{it}$. Since the n_i observations in the ith cluster are correlated, the joint density of the observations $\mathbf{y}_i=(y_{i1},\ldots,y_{in_i})'$ is

$$\Pr(\mathbf{y}_i|\beta,\mathbf{D})=\int\left\{\prod_{t=1}^{T}[\Phi(\mathbf{x}_{it}'\beta+\mathbf{w}_{it}'\mathbf{b}_i)]_{it}^{y}[1-\Phi(\mathbf{x}_{it}'\beta+\mathbf{w}_{it}'\mathbf{b}_i)]^{1-y_{it}}\right\}\times\mathcal{N}(\mathbf{b}_i|\mathbf{0},\mathbf{D})\,d\,\mathbf{b}_i$$

Under the assumption that the observations across individuals are independent, the likelihood function of the parameters $(\beta,\mathbf{D})$ is the product of $\Pr(\mathbf{y}_i|\beta,\mathbf{D})$. Although methods are now available to evaluate this integral under some special circumstances, it turns out that it is possible to circumvent the calculation of the likelihood function. The method relies on the approach that was introduced by Albert and Chib (1993).

To understand the *Albert and Chib* algorithm, consider the cross-section binary probit model in which we are given n random observations such that $\Pr(y_i=1)=\Phi(\mathbf{x}_i'\beta)$. An equivalent formulation of the model is in terms of latent variables $\mathbf{z}=(z_1,\ldots,z_n)$ where

$$z_i|\beta\sim\mathcal{N}(\mathbf{x}_i'\beta,1),\ y_i=I[z_i>0],$$

and I is the indicator function. Albert and Chib (1993) exploit this equivalence and propose that the latent variables $\{z_1,\ldots,z_n\}$, one for each observation, be included in the MCMC algorithm along with the regression parameter β. In other words, they suggest using MCMC methods to sample the joint posterior distribution

$$\pi(\beta,\mathbf{z}|\mathbf{y})\propto\pi(\beta)\prod_{i=1}^{N}\mathcal{N}(z_i|\mathbf{x}_i'\beta,1)\left\{I(z_i>0)^{y_i}+I(z_i<0)^{1-y_i}\right\}$$

where the term in braces in the probability of y_i given (β,z_i) and is one for $y_i=1$ when $z_i>0$ and is one for $y_i=0$ when $z_i<0$. The latter posterior density is sampled by a two-block Gibbs sampler composed of the full conditional distributions:

1. $\beta|\mathbf{y},\mathbf{z}$
2. $\mathbf{z}|\mathbf{y},\beta$.

Even though the parameter space has been enlarged, the introduction of the latent variables simplifies the problem considerably. The first conditional distribution, i.e., $\beta|\mathbf{y},\mathbf{z}$, is the same as the distribution $\beta|\mathbf{z}$ since knowledge of $\mathbf{z}$ means that $\mathbf{y}$ has no additional information for β. The distribution $\beta|\mathbf{z}$ is easy to derive since the response variable is continuous. The second conditional distribution, i.e., $\mathbf{z}|\mathbf{y},\beta$, factors into n distributions $z_i|y_i,\beta$ and is easily seen to be truncated normal given the value of y_i. Specifically, if $y_i=1$, then

$$z_i\sim\mathcal{TN}_{(0,\infty)}(\mathbf{x}_i'\beta,1)\tag{15.20}$$

a truncated normal distribution with support $(0,\infty)$, whereas if $y_i = 0$, then

$$z_i \sim \mathscr{TN}_{(-\infty,0]}(\mathbf{x}_i'\beta, 1) \tag{15.21}$$

a truncated normal distribution with support $(-\infty, 0)$. These truncated normal distributions are simulated by the method given in Sect. 1.4. For the case of (15.20), it reduces to

$$\mathbf{x}_i'\beta + \Phi^{-1}\left[\Phi(-\mathbf{x}_i'\beta) + U\left(1 - \Phi(-\mathbf{x}_i'\beta)\right)\right]$$

and for the case (15.21) to

$$\mathbf{x}_i'\beta + \Phi^{-1}\left[U\Phi(-\mathbf{x}_i'\beta)\right],$$

where U is a uniform random variable on (0,1). Hence, the algorithm proceeds through the simulation of β given the latent data and the simulation of the latent data given $(\mathbf{y}, \beta)$.

Given this framework, the approach for the panel probit model becomes transparent. For the ith cluster, we define the vector of latent variable

$$\mathbf{z}_i = \mathbf{X}_i\beta + \mathbf{W}_i\mathbf{b}_i + \varepsilon_i\,,\ \varepsilon_i \sim \mathscr{N}_{n_i}(\mathbf{0}, \mathbf{I}_{n_i})$$

and let

$$y_{it} = I[z_{it} > 0]$$

where $\mathbf{z}_i = (z_{i1}, \ldots, z_{in_i})'$, $\mathbf{W}_i$ is a set of variables whose effect $\mathbf{b}_i$ is assumed to heterogenous, $\mathbf{X}_i$ is a set of raw covariates or the matrix $\mathbf{W}_i\mathbf{A}_i$ or $(\mathbf{X}_{1i}\ \ \mathbf{W}_i\mathbf{A}_i)$ if the model is derived from a hierarchical specification in which the heterogeneity depends on cluster-specific covariates $\mathbf{A}_i$. The MCMC implementation in this set-up proceeds by including the $\{z_{it}\}$ in the sampling. Given the $\{z_{it}\}$ the sampling resembles the steps of the Gaussian–Gaussian algorithm with z_{it} playing the role of y_{it} and $\sigma^2\lambda_i^{-1}\mathbf{\Omega}_i = \mathbf{I}_{n_i}$. The sampling of z_{it} is done marginalized over $\{\mathbf{b}_i\}$ from the conditional distribution of $z_{it}|\mathbf{z}_{i(-t)}, y_{it}, \beta, \mathbf{D}$, where $\mathbf{z}_{i(-t)}$ is the vector $\mathbf{z}_i$ excluding z_{it}. It should be emphasized that the simulation of these distributions does not require the evaluation of the likelihood function.

Algorithm: Gaussian–Gaussian Panel Probit (Chib and Carlin (1999))

1. Sample

(a)

$$z_{it}|\mathbf{z}_{i(-t)}, y_{it}, \beta, \mathbf{D} \propto \mathscr{N}(\mu_{it}, v_{it})\left\{I(z_{it} < 0)^{1-y_{it}} + I(z_{it} > 0)^{y_{it}}\right\}$$
$$\mu_{it} = \mathrm{E}(z_{it}|\mathbf{z}_{i(-t)}, \beta, \mathbf{D})$$
$$v_{it} = \mathrm{Var}(z_{it}|\mathbf{z}_{i(-t)}, \beta, \mathbf{D})$$

(b)

$$\beta|\{z_{it}\}, \mathbf{D} \sim \mathscr{N}_k\left(\mathbf{B}(\mathbf{B}_0^{-1}\beta_0 + \sum_{i=1}^{N}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{z}_i), \mathbf{B}\right)$$

$$\mathbf{B} = (\mathbf{B}_0^{-1} + \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1};\ \mathbf{V}_i = \mathbf{I}_{n_i} + \mathbf{W}_i \mathbf{D} \mathbf{W}_i'$$

(c)

$$\mathbf{b}_i | \mathbf{y}, \beta, \mathbf{D} \sim \mathcal{N}_q \left(\mathbf{D}_i \mathbf{W}_i' (\mathbf{z}_i - \mathbf{X}_i \beta), \mathbf{D}_i \right), i \leq N$$

$$\mathbf{D}_i = (\mathbf{D}^{-1} + \mathbf{W}_i' \mathbf{W}_i)^{-1}$$

2. Sample

$$\mathbf{D}^{-1} | \mathbf{y}, \beta, \{\mathbf{b}_i\} \sim \mathcal{W}_q \{\rho_0 + N, \mathbf{R}\}$$

3. Goto 1

Because of this connection with the continuous case, the analysis of binary panel data may be extended in ways that parallel the developments in the previous section. For example, we can analyze binary data under the assumption that ε_i is multivariate-t and/or the assumption that the random effects distribution is student-t or a mixture of normals. We present the algorithm for the student–student binary response panel model without comment.

Algorithm: Student–Student Binary Panel

1. Sample

(a)

$$z_{it} | \mathbf{z}_{i(-t)}, y_{it}, \beta, \mathbf{D} \propto \mathcal{N}(\mu_{it}, v_{it}) \left\{ I(z_{it} < 0)^{1-y_{it}} + I(z_{it} > 0)^{y_{it}} \right\}$$

$$\mu_{it} = \mathrm{E}(z_{it} | \mathbf{z}_{i(-t)}, \beta, \mathbf{D}, \lambda_i)$$

$$v_{it} = \mathrm{Var}(z_{it} | \mathbf{z}_{i(-t)}, \beta, \mathbf{D}, \lambda_i)$$

(b)

$$\beta | \{z_{it}\}, \mathbf{D} \{\lambda_i\}, \{\eta_i\} \sim \mathcal{N}_k \left(\hat{\beta}, \mathbf{B} \right)$$

(c)

$$\mathbf{b}_i | \mathbf{z}_i, \beta, \mathbf{D}, \lambda_i, \eta_i \sim \mathcal{N}_q \left(\hat{\mathbf{b}}_i, \mathbf{D}_i \right)$$

2. Sample

(a)

$$\lambda_i | \mathbf{y}, \beta, \{\mathbf{b}_i\}, \sigma^2 \sim \mathcal{G} \left(\frac{\nu_G + n_i}{2}, \frac{\nu_G + \sigma^{-2} \mathbf{e}_i' \boldsymbol{\Omega}_i^{-1} \mathbf{e}_i}{2} \right), i \leq N$$

(b)

$$\eta_i | \mathbf{b}_i, \mathbf{D} \sim \mathcal{G} \left(\frac{\nu_G + q}{2}, \frac{\nu_G + \mathbf{b}_i' \mathbf{D}^{-1} \mathbf{b}_i}{2} \right), i \leq N$$

3. Sample

$$\mathbf{D}^{-1} | \{z_{it}\}, \beta, \{\mathbf{b}_i\}, \{\lambda_i\}, \{\eta_i\} \sim \mathcal{W}_q \{\rho_0 + N, \mathbf{R}\}$$

4. Goto 1

The fact that this, and other model variants for binary responses, are handled effortlessly is a testament to the flexibility and power of the Bayesian approach.

15.4 Other Outcome Types

15.4.1 Censored Outcomes

Given the discussion of the binary response models in the preceding section it should not be surprising that the Bayesian approach to censored data would proceed in much the same fashion. Consider then a Gaussian–Gaussian Tobit panel data model for the ith cluster:

$$\begin{aligned} z_i &= \mathbf{X}_i\beta + \mathbf{W}_i\mathbf{b}_i + \varepsilon_i \ , \ \varepsilon_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}) \\ \mathbf{b}_i &\sim \mathcal{N}_q(\mathbf{0}, \mathbf{D}) \\ (\beta, \sigma^2, \mathbf{D}) &\sim \pi \end{aligned}$$

where the observed outcomes are obtained as

$$y_{it} = \max\{z_{it}, 0\}$$

This model is fit along the lines of the Gaussian–Gaussian model by adopting the strategy of Chib (1992) wherein one simulates z_{it} for those observations that are censored from a truncated normal distribution, truncated to the interval $(-\infty, 0)$. In our description of the fitting method we let $\mathbf{y}_{i\mathbf{z}}$ be a $n_i \times 1$ vector with ith component y_{it} if that observation is not censored and z_{it} if it is censored. A new Step 1 is inserted in which the latent z_{it} are sampled conditioned on the remaining values of $\mathbf{y}_{i\mathbf{z}}$ in the ith cluster, which we denote by $y_{iz(-t)}$; then in Step 2 the only change is that instead of $\mathbf{y}_i$ we use $\mathbf{y}_{i\mathbf{z}}$; in Step 3 in the sampling of $\mathbf{b}_i$ we replace the vector $\mathbf{y}_i$ by the most current value of $\mathbf{y}_{i\mathbf{z}}$; Step 4 for the sampling of $\mathbf{D}^{-1}$ is unchanged; and in Step 5 dealing with the sampling of σ^2 we use $\mathbf{y}_{i\mathbf{z}}$ in place of $\mathbf{y}_i$ in the definition of δ.

Algorithm: Gaussian–Gaussian Tobit Panel

1. Sample

   (a)

$$\begin{aligned} z_{it}|y_{iz(-t)}, y_{it}, \beta, \sigma^2, \mathbf{D} &\propto \mathcal{N}(\mu_{it}, v_{it})\, I(z_{it} < 0) \quad \text{if } y_{it} = 0 \\ \mu_{it} &= \mathrm{E}(z_{it}|y_{iz(-t)}, \beta, \sigma^2, \mathbf{D}) \\ v_{it} &= \mathrm{Var}(z_{it}|y_{iz(-t)}, \beta, \sigma^2, \mathbf{D}) \end{aligned}$$

   (b)

$$\beta|\mathbf{y}_z, \sigma^2, \mathbf{D} \sim \mathcal{N}_k\left(\hat{\beta}, \mathbf{B}\right)$$

(c)
$$\mathbf{b}_i|\mathbf{y}_z,\beta,\sigma^2,\mathbf{D}\sim\mathscr{N}_q\left(\hat{\mathbf{b}}_i,\mathbf{D}_i\right),i\leq N$$

2. Sample
$$\mathbf{D}^{-1}|\mathbf{y}_z,\beta,\{\mathbf{b}_i\},\sigma^2\sim\mathscr{W}_q\{\rho_0+N,\mathbf{R}\}$$

3. Sample
$$\sigma^2|\mathbf{y}_z,\beta,\{\mathbf{b}_i\},\mathbf{D}\sim\mathscr{IG}\left(\frac{\nu_0+\sum n_i}{2},\frac{\delta_0+\delta}{2}\right)$$

4. Goto 1

Just as in the case of continuous and binary outcomes, this algorithm is easily modified to allow the random effects have a student-t or a mixture of normals distribution and to allow the observation errors be student-t. Analysis of any of these models is quite difficult from the frequentist perspective.

15.4.2 Count Responses

Bayesian methods are also effectively applied to panel data in which the responses are counts. A framework for fitting such models under the assumption that the distribution of the counts, given the random effects, is Poisson is developed by Chib, Greenberg and Winklemnann (1998). To describe the set-up, for the *i*th cluster

$$\begin{aligned}y_{it}|\beta,\mathbf{b}_i&\sim\text{Poisson}(\lambda_{it})\\ \ln(\lambda_{it})&=\ln\tau_{it}+\mathbf{x}_{it}'\beta+\mathbf{w}_{it}'\mathbf{b}_i\end{aligned}$$

where the covariate vectors $\mathbf{x}_{it}'$ and $\mathbf{w}_{it}'$ are the tth row of the matrices $\mathbf{X}_i$ and $\mathbf{W}_i$, respectively, and $\mathbf{X}_i$ are the raw covariates or the matrix $\mathbf{W}_i\mathbf{A}_i$ or $(\mathbf{X}_{1i}\ \mathbf{W}_i\mathbf{A}_i)$ if the model is derived from a hierarchical specification in which the heterogeneity depends on cluster-specific covariates $\mathbf{A}_i$. The quantity τ_{it} which is one if each count is measured over the same interval of time. This specification of the model produces the likelihood function

$$f(\mathbf{y}|\beta,\mathbf{D})=\prod_{i=1}^{n}\int\mathscr{N}(\mathbf{b}_i|\mathbf{0},\mathbf{D})p(\mathbf{y}_i|\beta,\mathbf{b}_i)d\mathbf{b}_i \tag{15.22}$$

where

$$p(\mathbf{y}_i|\beta,\mathbf{b}_i)=\prod_{t=0}^{n_i}\frac{\lambda_{it}^{y_{it}}\exp(-\lambda_{it})}{y_{it}!} \tag{15.23}$$

is the product of the Poisson mass function with mean λ_{it}.

The interesting aspect of the MCMC algorithm in this case is the sampling of both β and $\{\mathbf{b}_i\}$ by tailored M–H steps. This is because the full conditional distributions in this model do not belong to any known family of distributions. At each step of the algorithm, there are $n+1$ M–H steps. It may appear that the computational burden is high when n is large. This turns out not to be case.

Algorithm: Panel Poisson Chib, Greenberg and Winklemnann (1998))

1. Calculate the parameters $(\mathbf{m}_0, \mathbf{V}_0)$ as the mode and inverse of the negative Hessian of
$$\log \mathcal{N}_k(\beta|\beta_0, \mathbf{B}_0) + \sum_{i=1}^{N} \log p(\mathbf{y}_i|\beta, \mathbf{b}_i)$$
propose $\beta' \sim \mathcal{T}(\beta|\mathbf{m}_0, \mathbf{V}_0, \nu)$ (the multivariate-t density) and move to β' with probability
$$\min\left\{\frac{\prod_{i=1}^N p(\mathbf{y}_i|\beta', \mathbf{b}_i)\mathcal{N}_k(\beta'|\mathbf{0}, \mathbf{B})}{\prod_{i=1}^N p(\mathbf{y}_i|\beta, \mathbf{b}_i)\mathcal{N}_k(\beta|\mathbf{0}, \mathbf{B})} \frac{\mathcal{T}(\beta|\mathbf{m}_0, \mathbf{V}_0, \nu)}{\mathcal{T}(\beta'|\mathbf{m}_0, \mathbf{V}_0, \nu)}, 1\right\}$$
2. Calculate the parameters $(\mathbf{m}_i, \mathbf{V}_i)$ as the mode and inverse of the negative Hessian of
$$\log \mathcal{N}_q(\mathbf{b}_i|\mathbf{0}, \mathbf{D}) + \log p(\mathbf{y}_i|\beta, \mathbf{b}_i)$$
propose $\mathbf{b}_i' \sim \mathcal{T}(\mathbf{b}_i|\mathbf{m}_i, \mathbf{V}_i, \nu)$ and move to $\mathbf{b}_i'$ with probability
$$\min\left\{\frac{p(\mathbf{y}_i|\beta, \mathbf{b}_i')\mathcal{N}_q(\mathbf{b}_i'|\mathbf{0}, \mathbf{D})}{p(\mathbf{y}_i|\beta, \mathbf{b}_i)\mathcal{N}_q(\mathbf{b}_i|\mathbf{0}, \mathbf{D})} \frac{\mathcal{T}(\mathbf{b}_i|\mathbf{m}_i, \mathbf{V}_i, \nu)}{\mathcal{T}(\mathbf{b}_i'|\mathbf{m}_i, \mathbf{V}_i, \nu)}, 1\right\}$$
3. Sample
$$\mathbf{D}^{-1}|\mathbf{y}, \beta, \{\mathbf{b}_i\}, \sigma^2 \sim \mathcal{W}_q\{\rho_0 + N, \mathbf{R}\}$$

15.4.3 Multinomial Responses

Multinomial panel responses arise in several different areas and the fitting of this model when the link function is assumed to be multinomial logit is exactly the same as the algorithm for count responses. The only difference is that instead of the Poisson link function we now have the multinomial logit link. Let y_{it} be a multinomial random variable taking values $\{0, 1, \ldots, J\}$ and assume that
$$\Pr(y_{it} = j|\beta, \mathbf{b}_i) = \frac{\exp\left(\alpha_j + \mathbf{x}_{itj}'\beta + \mathbf{w}_{itj}'\mathbf{b}_i\right)}{\sum_{l=0}^J \exp\left(\alpha_l + \mathbf{x}_{itl}'\beta + \mathbf{w}_{itl}'\mathbf{b}_i\right)}$$
where for identifiability α_0 is set equal to zero. The joint probability of the outcomes in the ith cluster, conditioned on the $\mathbf{b}_i$, is now given by
$$p(\mathbf{y}_i|\beta, \mathbf{b}_i) = \prod_{t=1}^{n_i} \Pr(y_{it} = j_t|\beta, \mathbf{b}_i) \tag{15.24}$$
where j_t is the observed outcome at time t. The structure of the problem is seen to be identical to that in the count case and the preceding algorithm applies directly to this problem by replacing the mass function in (15.23) with the mass function in (15.24).

Chiang, Chib and Narasimhan (1999) develop an interesting variant of this model in which the possible values that y_{it} can take is not the same across clusters. Such a situation arises when the multinomial outcomes are choices made by a subject (for example choice of transportation mode or choice of brand of a product) and where the assumption that the choice set is the same across subjects is too strong and must be relaxed. The model discussed in the paper only appears to be fittable by Bayesian methods. The paper includes a detailed example.

15.5 Binary Endogenous Regressor

In many applied studies, one is interested in the effect of a given (binary) covariate on the response but under the complication that the binary covariate is not sequentially randomly assigned. In other words, the assumption of sequential exogeneity is violated. This problem has not been extensively studied in the literature but interestingly it is possible to develop a Bayesian approach to inference that in many ways is quite straightforward. For concreteness, suppose that in the context of the model in (15.2) the last covariate in $\mathbf{x}_{1it}$ (namely x_{12it}) is the covariate of interest and the model is given by

$$y_{it} = \mathbf{x}'_{11it}\beta_{11} + x_{12it}\beta_{12} + \mathbf{w}'_{it}\mathbf{c}_{2i} + e_{it}$$

where $\mathbf{x}_{1it} = (\mathbf{x}_{11it}, x_{12it})$ and $\mathbf{x}_{11it} : k_{11} \times 1$. Assume that the covariates $\mathbf{x}_{11it}$ and $\mathbf{w}_{it} : q \times 1$ satisfy the assumption of sequential exogeneity but that x_{12it} does not. Now let $z_{it} : k_z \times 1$ be time-varying instruments and suppose that the model generating the endogenous covariate is

$$x_{12it} = I\left(\mathbf{x}'_{11it}\gamma + \mathbf{w}'_{it}\mathbf{d}_{2i} + \mathbf{z}_{it}\delta + u_{it} > 0\right)$$

where

$$\begin{pmatrix} e_{it} \\ u_{it} \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{\Omega} = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & 1 \end{pmatrix}\right)$$

and $\omega_{12} \neq 0$. Letting $x^*_{12it} = \mathbf{x}'_{11it}\gamma + \mathbf{w}'_{it}\mathbf{d}_{3i} + \mathbf{z}_{it}\delta + u_i$, the model is reexpressed as

$$\underbrace{\begin{pmatrix} y_{it} \\ x^*_{12it} \end{pmatrix}}_{\mathbf{y}^*_{it}} = \underbrace{\begin{pmatrix} \mathbf{x}'_{11it} & x_{12it} & \mathbf{0}' & \mathbf{0}' \\ \mathbf{0}' & 0 & \mathbf{x}'_{11it} & z_{it} \end{pmatrix}}_{\mathbf{X}_{1it}} \underbrace{\begin{pmatrix} \beta_{11} \\ \beta_{12} \\ \gamma \\ \delta \end{pmatrix}}_{\beta_1} + \underbrace{\begin{pmatrix} \mathbf{w}'_{it} & \mathbf{0}' \\ \mathbf{0}' & \mathbf{w}'_{it} \end{pmatrix}}_{\mathbf{W}_{it}} \underbrace{\begin{pmatrix} \mathbf{c}_{2i} \\ \mathbf{d}_{2i} \end{pmatrix}}_{\beta_{2i}} + \underbrace{\begin{pmatrix} e_{it} \\ u_{it} \end{pmatrix}}_{\varepsilon_{it}}$$

or as

$$\mathbf{y}^*_{it} = \mathbf{X}_{1it}\beta_1 + \mathbf{W}_{it}\beta_{2i} + \varepsilon_{it}$$

where β_1 is $k_1 \times 1$ with $k_1 = 2k_{11} + 1 + k_z$ and β_{2i} is $2q \times 1$. If we assume that β_{2i} as before is modeled in terms of covariates $\mathbf{a}_i : r \times 1$ as

$$\begin{pmatrix} \mathbf{c}_{2i} \\ \mathbf{d}_{2i} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_q \otimes \mathbf{a}_i' & \mathbf{0}' \\ \mathbf{0}' & \mathbf{I}_q \otimes \mathbf{a}_i' \end{pmatrix} \begin{pmatrix} \beta_{21} \\ \beta_{22} \end{pmatrix} + \begin{pmatrix} \mathbf{b}_{1i} \\ \mathbf{b}_{2i} \end{pmatrix}$$

or compactly as

$$\beta_{2i} = \mathbf{A}_i \beta_2 + \mathbf{b}_i$$

where $\beta_2 : k_2 \times 1$ and $k_2 = 2qr$, then we can rewrite the outcome vector for subject i at time t as

$$\mathbf{y}_{it}^* = \mathbf{X}_{it}\beta + \mathbf{W}_{it}\mathbf{b}_i + \varepsilon_{it}$$

where

$$\mathbf{X}_{it} = (\mathbf{X}_{1it}, \mathbf{A}_i \mathbf{W}_{it})$$

$\beta = (\beta_1, \beta_2) : k \times 1$, and $k = k_1 + k_2$. This is similar to the models that we have dealt with except that this is a system of two equations for each (i,t) with the second component of the outcome being latent. For the ith cluster the preceding model (in conjunction with the standard assumptions about $\mathbf{b}_i$) is written as

$$\mathbf{y}_i^* = \mathbf{X}_i\beta + \mathbf{W}_i\mathbf{b}_i + \varepsilon_i \ , \ \varepsilon_i | \lambda_i, \mathbf{\Omega} \sim \mathcal{N}_{2n_i}(\mathbf{0}, \lambda_i^{-1} \{I_{n_i} \otimes \mathbf{\Omega}\})$$
$$\mathbf{b}_i | \mathbf{D} \sim \mathcal{N}_{2q}(\mathbf{0}, \mathbf{D})$$

This model is fit along the lines of the binary panel by simulating $\{x_{12it}^*\}_{t=1}^{n_i}$ in $\mathbf{y}_i^*$ (these appear in rows 2, 4, 6, etc. in the vector y_i^*) from appropriate truncated normal distributions, according to the device of Albert and Chib (1993), marginalized over $\mathbf{b}_i$. In our description of the fitting method given below it is to be understood that $\mathbf{y}_i^*$ contains the most recently simulated values of $\{x_{12it}^*\}_{t=1}^{n_i}$. A new step is the sampling of $(\omega_{11}, \omega_{12})$. The best way of working with these parameters is to reparameterize them to (σ^2, ω_{12}) where $\sigma^2 = \omega_{11} - \omega_{12}^2$ and then assuming that prior information on the transformed parameters is represented by the conditionally conjugate distribution

$$\pi(\sigma^2, \omega_{12}) = \mathcal{IG}\left(\sigma^2 | \frac{\nu_0}{2}, \frac{\delta_0}{2}\right) \mathcal{N}\left(\omega_{12} | m_0, \sigma^2 M_0\right) \tag{15.25}$$

Now conditioned on $\{x_{12it}^*\}_{i,t}$ and $\{\mathbf{b}_i\}$ it follows that

$$\widetilde{y}_{it} = \omega_{12} u_{it} + v_{it} \tag{15.26}$$

where

$$\widetilde{y}_{it} = y_{it} - \mathbf{x}_{11it}'\beta_{11} - x_{12it}\beta_{12} - \mathbf{w}_{it}'\mathbf{c}_{2i} \ ,$$
$$u_{it} = x_{12it}^* - \mathbf{x}_{11it}'\gamma + \mathbf{w}_{it}'\mathbf{d}_{2i} + z_{it}\delta$$

and

$$v_{it} \sim N(0, \sigma^2)$$

The prior in (15.25) and the sampling model in (15.26) when combined by Bayes theorem produce an updated distribution of (σ^2, ω_{12}) that is sampled in one block. To see the details, we express the model in (15.26) for all $M = \sum_{i=1}^{N} T_i$ observations as

$$\widetilde{\mathbf{y}} = \omega_{12}\mathbf{u} + \mathbf{v}$$

where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_M)$. By simple calculations it is seen that the updated distribution of σ^2 marginalized over ω_{12} is

$$\mathcal{IG}\left(\sigma^2|\frac{\nu_0 + M}{2}, \frac{\delta_0 + (\widetilde{\mathbf{y}} - \mathbf{u}m_0)'(I_M + \mathbf{u}M_0\mathbf{u}')^{-1}(\widetilde{\mathbf{y}} - \mathbf{u}m_0)}{2}\right)$$

while that of ω_{12} conditioned on σ^2 is

$$\mathcal{N}\left(\omega_{12}|, W\sigma^{-2}\left(M_0 m_0 + \mathbf{u}'\widetilde{\mathbf{y}}\right), W = \sigma^2\left(M_0 + \mathbf{u}'\mathbf{u}\right)^{-1}\right)$$

Algorithm: Gaussian–Gaussian Binary Endogenous Panel

1. Sample

   (a)

$$\begin{aligned} x^*_{12it}|\left(y^*_i \backslash x^*_{12it}\right), x_{12it}, \beta, \mathbf{D}, \mathbf{\Omega} &\sim\propto \mathcal{N}\left(\mu_{it}, v_{it}\right) \\ &\left\{I\left(x^*_{12it} < 0\right)^{1-x_{12it}} + I\left(x_{12it} > 0\right)^{x_{12it}}\right\} \\ \mu_{it} &= \mathrm{E}(x^*_{12it}|\left(y^*_i \backslash x^*_{12it}\right), x_{12it}, \beta, \mathbf{D}, \mathbf{\Omega}) \\ v_{it} &= \mathrm{Var}(x^*_{12it}|\left(y^*_i \backslash x^*_{12it}\right), x_{12it}, \beta, \mathbf{D}, \mathbf{\Omega}) \end{aligned}$$

   (b)

$$\beta|\mathbf{y}^*, \mathbf{\Omega}, \mathbf{D} \sim \mathcal{N}_k\left(\hat{\beta}, \mathbf{B}\right)$$

   (c)

$$\mathbf{b}_i|\mathbf{y}^*, \beta, \mathbf{D}, \mathbf{\Omega} \sim \mathcal{N}_q\left(\hat{\mathbf{b}}_i, \mathbf{D}_i\right), i \leq N$$

2. Sample

$$\mathbf{D}^{-1}|\mathbf{y}^*, \beta, \{\mathbf{b}_i\}, \sigma^2 \sim \mathcal{W}_q\left\{\rho_0 + N, \mathbf{R}\right\}$$

3. Sample

   (a)

$$\sigma^2|\mathbf{y}^*, \beta, \{\mathbf{b}_i\}, \mathbf{D} \sim \mathcal{IG}\left(\sigma^2|\frac{\nu_0 + M}{2}, \frac{\delta_0 + (\widetilde{\mathbf{y}} - \mathbf{u}m_0)'(I_M + \mathbf{u}M_0\mathbf{u}')^{-1}(\widetilde{\mathbf{y}} - \mathbf{u}m_0)}{2}\right)$$

(b)

$$\omega_{12}|\mathbf{y}^*, \beta, \{\mathbf{b}_i\}, \sigma^2, \mathbf{D} \sim$$
$$\mathcal{N}\left(\omega_{12}|, W\sigma^{-2}\left(M_0 m_0 + \mathbf{u}'\widetilde{\mathbf{y}}\right), W = \sigma^2\left(M_0 + \mathbf{u}'\mathbf{u}\right)^{-1}\right)$$

4. Goto 1

15.6 Informative Missingness

It is possible to develop a range of panel data models in which the outcome on a given subject at time t is potentially missing. Each individual at time t supplies two observations: c_{it} and y_{it}. The variable c_{it} is binary and takes the value 1 in which case y_{it} is observed or the value 0 in which case the observation y_{it} is missing. The two random variables are correlated due to the presence of common unobserved random variables. The missingness mechanism is thus non-ignorable. To describe the basic components of such a model, suppose y_{it} is the outcome (which could be continuous, discrete, or censored) and c_{it} is an indicator variable of non-missingness. As an example suppose that the variable c_{it} is one if the individual is working and 0 otherwise and y_{it} is a continuous variable indicating the person's wage. Thus, the variable y_{it} is observed when c_{it} is one; otherwise the variable y_{it} is missing. Let c_{it}^* denote a continuous random variable that is marginally generated as

$$c_{it}^* = \mathbf{x}_{it}'\gamma_i + z_{it}\delta_i + u_i$$

and let

$$c_{it} = I\left(c_{it}^* > 0\right)$$

where γ_i and δ_i are subject-specific coefficients and z_{it} is an additional covariate (the instrument). For simplicity we are assuming that the effect of each covariate is subject-specific although this can be relaxed, much as we have done in the models discussed previously. Also suppose that the outcome y_{it} (under the assumption that it is continuous) is marginally generated as

$$y_{it} = \mathbf{x}_{it}'\alpha_i + \varepsilon_{it}$$

where

$$\begin{pmatrix} \varepsilon_{it} \\ u_{it} \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{\Omega} = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & 1 \end{pmatrix}\right)$$

To complete the model, we specify the distribution of the heterogenous coefficients with a hierarchical prior. Let $\beta_i = (\alpha_i', \gamma_i', \delta_i)'$ and assume that

$$\beta_i|\beta, \mathbf{D} \sim \mathcal{N}\left(\mathbf{A}_i\beta, \mathbf{D}\right)$$

where $\mathbf{D}$ is a full matrix. Under this latter specification, the two components of the model are tied together not just by correlation amongst the errors but also by the

dependence between α_i and (γ_i, δ_i) as measured by the off-diagonal blocks D_{12} and D_{13} of D. It is also assumed that the covariates x_{it} and z_{it} are observable even when y_{it} is missing (ie., when $c_{it} = 0$).

We mention that a variant of this model is considered by Chib, Seetharaman and Strijnev (2004). In that model y_{it} is multinomial indicating choice amongst a given set of brands in a particular category (say cola) and c_{it} is a variable that indicates whether purchase into the category occurs at shopping visit t; if the individual does not purchase in the category then the brand-choice outcome is missing. They describe the Bayesian MCMC fitting of the model and apply the model and the algorithm to a scanner panel data set.

15.7 Prediction

In some problems one is interested in predicting one or more post-sample observations on a given individual. Specifically, for an individual in the sample, we are interested in making inferences about the set of observations

$$\mathbf{y}_{if} = (y_{in_i+1}, \ldots, y_{in_i+s})$$

given sample data and a particular hierarchical Bayesian model. In the Bayesian context, the problem of prediction is solved by the calculation of the predictive density

$$f(\mathbf{y}_{if}|\mathbf{y}) = \int p(\mathbf{y}_{if}|\mathbf{y}, \delta_i, \boldsymbol{\theta})\pi(\delta_i, \boldsymbol{\theta}|\mathbf{y})d\delta_i d\boldsymbol{\theta}$$

where δ_i denotes the set of cluster-specific unobserved random-variables (such as $\mathbf{z}_i$ in binary and censored response models and the random effects $\mathbf{b}_i$) and $\boldsymbol{\theta}$ denote the entire set of parameters. The predictive density is the density of $\mathbf{y}_{if}$ marginalized over $(\delta_i, \boldsymbol{\theta})$ with respect to the posterior distribution of (δ_i, θ).

This predictive density is summarized in the same way that we summarized the posterior density of the parameters–by sampling it. Sampling of the predictive density is conducted by the *method of composition*. According to the method of composition, if $f(\mathbf{y}) = \int f(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})\mathrm{d}\mathbf{x}$, and $\mathbf{x}^{(g)}$ is a draw from $\pi(\mathbf{x})$, then $\mathbf{y}^{(g)}$ drawn from $f(\mathbf{y}|\mathbf{x}^{(g)})$ is a draw from $f(\mathbf{y})$. Thus, a draw from the marginal is obtained simply by sampling the conditional density $f(\mathbf{y}|\mathbf{x})$ for each value drawn from $\pi(\mathbf{x})$.

The method of composition leads to an easily implementable procedure for calculating the predictive density in every panel data model that we have considered. For example in the Gaussian-Gaussian model, given $(\beta^{(g)}, \sigma^{2(g)}, \mathbf{b}_i^{(g)})$, the gth MCMC draw on $(\beta, \sigma^2, \mathbf{b}_i)$, the gth draw from the predictive density is obtained by drawing

$$\varepsilon_{it}^{(g)} \sim \mathcal{N}(0, \sigma^{2(g)}) \ , \ t = n_i + 1, \ldots, n_i + s$$

and setting

$$y_{it}^{(g)} = \mathbf{x}_{it}'\beta^{(g)} + \mathbf{w}_{it}'\mathbf{b}_i^{(g)} + \varepsilon_{it}^{(g)} \ , \ t = n_i + 1, \ldots, n_i + s$$

The resulting sample of draws are summarized in terms of moments, quantiles and density plots.

15.8 Residual Analysis

One approach to Bayesian residual analysis relies on the idea of "realized errors" introduced by Zellner (1975) and studied more recently by Chaloner and Brant (1988) and Albert and Chib (1995). The idea is to compute the posterior distribution of the error and define a residual to be outlying if the posterior distribution is concentrated on large values.

Consider for simplicity the Gaussian–Gaussian model for continuous responses. In that case, the error conditioned on y_{it} is given by

$$\varepsilon_{it} = y_{it} - \mathbf{x}_{it}'\beta - \mathbf{w}_{it}'\mathbf{b}_i$$

and, therefore, the posterior distribution of ε_{it} is determined by the posterior distribution of β and $\mathbf{b}_i$. To obtain this posterior distribution, at each iteration of the sampling, we compute the value

$$\varepsilon_{it}^{(g)} = y_{it} - \mathbf{x}_{it}'\beta^{(g)} - \mathbf{w}_{it}'\mathbf{b}_i^{(g)}$$

where $\{\beta^{(g)}, \mathbf{b}_i^{(g)}\}$ are the gth sampled values. Then, the collection of values $\{\varepsilon_{it}^{(g)}\}$ constitutes a sample from the posterior distribution $\pi(\varepsilon_{it}|\mathbf{y})$. There are various ways to summarize this posterior distribution in order to find outlying observations. One possibility is to compute the posterior probability

$$\Pr\left(\left|\frac{\varepsilon_{it}}{\sigma}\right| > k|\mathbf{y}\right)$$

where k is 2 or 3, and compare the posterior probability (computed from the simulated draws $\varepsilon_{it}^{(g)}/\sigma^{(g)}$) with the prior probability that the standardized residual is bigger that k in absolute value. The observation is classified as on outlier if the ratio of the posterior probability to the prior probability is large. Interestingly, similar ideas are used in panel probit models as discussed by Albert and Chib (1995).

15.9 Model Comparisons

Posterior simulation by MCMC methods does not require knowledge of the normalizing constant of the posterior density. Nonetheless, if we are interested in comparing alternative models, then knowledge of the normalizing constant is essential. This is because the standard and formal Bayesian approach for comparing models is via *Bayes factors*, or ratios of *marginal likelihoods*. The marginal likelihood of a particular model is the normalizing constant of the posterior density and is defined as

$$m(\mathbf{y}|\mathscr{M}) = \int p(\mathbf{y}|\mathscr{M}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathscr{M})\mathrm{d}\boldsymbol{\theta}\,, \tag{15.27}$$

the integral of the likelihood function with respect to the prior density. If we have two models $\mathscr{M}_k$ and $\mathscr{M}_l$, then the Bayes factor is the ratio

$$\mathrm{B}_{kl} = \frac{m(\mathbf{y}|\mathscr{M}_k)}{m(\mathbf{y}|\mathscr{M}_l)} \,. \tag{15.28}$$

Computation of the marginal likelihood is, therefore, of some importance in Bayesian statistics (DiCicio, Kass, Raftery and Wasserman (1997), Chen and Shao (1998), Roberts (2001)). Unfortunately, because MCMC methods deliver draws from the posterior density, and the marginal likelihood is the integral with respect to the prior, the MCMC output cannot be used directly to average the likelihood. To deal with this problem, a number of methods have appeared in the literature. One simple and widely applicable method is due to Chib (1995) which we briefly explain as follows.

Begin by noting that $m(\mathbf{y})$ by virtue of being the normalizing constant of the posterior density can be expressed as

$$m(\mathbf{y}|\mathscr{M}) = \frac{p(\mathbf{y}|\mathscr{M},\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*|\mathscr{M})}{\pi(\boldsymbol{\theta}^*|\mathscr{M},\mathbf{y})} \,, \tag{15.29}$$

for any given point $\boldsymbol{\theta}^*$ (generally taken to be a high density point such as the posterior mean). Thus, provided we have an estimate $\hat{\pi}(\boldsymbol{\theta}^*|\mathscr{M},\mathbf{y})$ of the posterior ordinate, the marginal likelihood is estimated on the log scale as

$$\log m(\mathbf{y}|\mathscr{M}) = \log p(\mathbf{y}|\mathscr{M},\theta^*) + \log \pi(\boldsymbol{\theta}^*|\mathscr{M}) - \log \hat{\pi}(\theta^*|\mathscr{M},\mathbf{y}) \,. \tag{15.30}$$

In the context of both single and multiple block M–H chains, good estimates of the posterior ordinate are available. For example, when the MCMC simulation is run with B blocks, to estimate the posterior ordinate we employ the marginal-conditional decomposition

$$\pi(\boldsymbol{\theta}^*|\mathscr{M},\mathbf{y}) = \pi(\boldsymbol{\theta}_1^*|\mathscr{M},\mathbf{y}) \times \ldots \times \pi(\boldsymbol{\theta}_i^*|\mathscr{M},\mathbf{y},\boldsymbol{\psi}_{i-1}^*) \times \ldots \times \pi(\boldsymbol{\theta}_B^*|\mathscr{M},\mathbf{y},\psi_{B-1}^*) \,, \tag{15.31}$$

where on letting $\boldsymbol{\psi}_i = (\boldsymbol{\theta}_1,\ldots,\theta_i)$ and $\boldsymbol{\psi}^i = (\boldsymbol{\theta}_i,\ldots,\theta_B)$ denote the list of blocks upto i and the set of blocks from i to B, respectively, and $\mathbf{z}$ denoting the latent data, and dropping the model index for notational convenience, the typical term is of the form

$$\pi(\boldsymbol{\theta}_i^*|\mathbf{y},\boldsymbol{\psi}_{i-1}^*) = \int \pi(\boldsymbol{\theta}_i^*|\mathbf{y},\ \psi_{i-1}^*,\boldsymbol{\psi}^{i+1},\mathbf{z})\,\pi(\boldsymbol{\psi}^{i+1},\mathbf{z}|\,\mathbf{y},\boldsymbol{\psi}_{i-1}^*)\mathrm{d}\boldsymbol{\psi}^{i+1}\mathrm{d}\mathbf{z}$$

This is the *reduced conditional ordinate*. It is important to bear in mind that in finding the reduced conditional ordinate one must integrate only over $(\boldsymbol{\psi}^{i+1},\mathbf{z})$ and that the integrating measure is conditioned on $\boldsymbol{\psi}_{i-1}^*$.

Consider first the case where the normalizing constant of each full conditional density is known. Then, the first term of (15.31) is estimated by the Rao–Blackwell method. To estimate the typical reduced conditional ordinate, one conducts a MCMC run consisting of the full conditional distributions

$$\left\{\pi(\boldsymbol{\theta}_i|\mathbf{y},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^{i+1},\mathbf{z});\ldots;\pi(\theta_B|\mathbf{y},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\theta}_i,\ldots,\boldsymbol{\theta}_{B-1},\mathbf{z})\,;\pi(\mathbf{z}|\mathbf{y},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^i)\right\} \tag{15.32}$$

where the blocks in $\boldsymbol{\psi}_{i-1}$ are set equal to $\boldsymbol{\psi}^*_{i-1}$. By MCMC theory, the draws on $(\boldsymbol{\psi}^{i+1},\mathbf{z})$ from this run are from the distribution $\pi(\boldsymbol{\psi}^{i+1},\mathbf{z}|\mathbf{y},\boldsymbol{\psi}^*_{i-1})$ and so the reduced conditional ordinate is estimated as the average

$$\hat{\pi}(\boldsymbol{\theta}^*_i|\,\mathbf{y},\boldsymbol{\psi}^*_{i-1}) = M^{-1}\sum_{j=1}^{M}\pi(\boldsymbol{\theta}^*_i|\mathbf{y},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^{i+1,(j)},\mathbf{z}^{(j)})$$

over the simulated values of $\boldsymbol{\psi}^{i+1}$ and $\mathbf{z}$ from the reduced run. Each subsequent reduced conditional ordinate that appears in the decomposition (15.31) is estimated in the same way though, conveniently, with fewer and fewer distributions appearing in the reduced runs. Given the marginal and reduced conditional ordinates, the marginal likelihood on the log scale is available as

$$\log\hat{m}(\mathbf{y}|\mathscr{M}) = \log p(\mathbf{y}|\mathscr{M},\theta^*) + \log\pi(\boldsymbol{\theta}^*|\mathscr{M}) - \sum_{i=1}^{B}\log\hat{\pi}(\,\theta^*_i|\mathscr{M},\mathbf{y},\boldsymbol{\psi}^*_{i-1}) \tag{15.33}$$

where $p(\mathbf{y}|\mathscr{M},\theta^*)$ is the density of the data marginalized over the latent data $\mathbf{z}$.

Consider next the case where the normalizing constant of one or more of the full conditional densities is not known. In that case, the posterior ordinate is estimated by a modified method developed by Chib and Jeliazkov (2001). If sampling is conducted in one block by the M–H algorithm, then it can be shown that the posterior ordinate is given by

$$\pi(\boldsymbol{\theta}^*|\mathbf{y}) = \frac{E_1\{\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^*|\mathbf{y})q(\,\theta,\boldsymbol{\theta}^*|\mathbf{y})\}}{E_2\{\alpha(\boldsymbol{\theta}^*,\boldsymbol{\theta}|\mathbf{y})\}}$$

where the numerator expectation E_1 is with respect to the distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ and the denominator expectation E_2 is with respect to the proposal density of $\boldsymbol{\theta}$ conditioned on $\boldsymbol{\theta}^*$, $q(\boldsymbol{\theta}^*,\boldsymbol{\theta}|\mathbf{y})$, and $\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^*|\mathbf{y})$ is the probability of move in the M–H step. This leads to the simulation consistent estimate

$$\hat{\pi}(\boldsymbol{\theta}^*|\mathbf{y}) = \frac{M^{-1}\sum_{g=1}^{M}\alpha(\boldsymbol{\theta}^{(g)},\boldsymbol{\theta}^*|\mathbf{y})q(\boldsymbol{\theta}^{(g)},\boldsymbol{\theta}^*|\mathbf{y})}{J^{-1}\sum_{j=1}^{M}\alpha(\boldsymbol{\theta}^*,\boldsymbol{\theta}^{(j)}|\mathbf{y})}\,, \tag{15.34}$$

where $\{\boldsymbol{\theta}^{(g)}\}$ are the given draws from the posterior distribution while the draws $\boldsymbol{\theta}^{(j)}$ in the denominator are from $q(\boldsymbol{\theta}^*,\boldsymbol{\theta}|\mathbf{y})$, given the fixed value $\boldsymbol{\theta}^*$.

In general, when sampling is done with B blocks, the typical reduced conditional ordinate is given by

$$\pi(\boldsymbol{\theta}^*_i|\mathbf{y},\boldsymbol{\theta}^*_1,\ldots,\boldsymbol{\theta}^*_{i-1}) = \frac{E_1\left\{\alpha(\boldsymbol{\theta}_i,\boldsymbol{\theta}^*_i|\mathbf{y},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^{i+1})q_i(\boldsymbol{\theta}_i,\boldsymbol{\theta}^*_i|\mathbf{y},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^{i+1})\right\}}{E_2\left\{\alpha(\boldsymbol{\theta}^*_i,\boldsymbol{\theta}_i|\mathbf{y},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^{i+1})\right\}} \tag{15.35}$$

where E_1 is the expectation with respect to $\pi(\boldsymbol{\psi}^{i+1}|\mathbf{y},\boldsymbol{\psi}^*_{i-1})$ and E_2 that with respect to the product measure $\pi(\boldsymbol{\psi}^{i+1}|\mathbf{y},\boldsymbol{\psi}^*_i)q_i(\boldsymbol{\theta}^*_i,\boldsymbol{\theta}_i|\mathbf{y},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^{i+1})$. The quantity $\alpha(\boldsymbol{\theta}_i,\boldsymbol{\theta}^*_i|\mathbf{y},\boldsymbol{\psi}^*_{i-1},\boldsymbol{\psi}^{i+1})$ is the usual *conditional* M–H probability of move. The two expectations are estimated from the output of the reduced runs in an obvious way.

15.9.1 Gaussian–Gaussian Model

As an example of the calculation of the marginal likelihood consider the calculation of the posterior ordinate for the Gaussian–Gaussian continuous response model. The ordinate is written as

$$\pi(\mathbf{D}^{-1*},\sigma^{2*},\beta^*|\mathbf{y}) = \pi(\mathbf{D}^{-1*}|\mathbf{y})\pi(\sigma^{2*}|\mathbf{y},\mathbf{D}^*)\pi(\beta^*|\mathbf{y},\mathbf{D}^*,\sigma^{2*}) ,$$

where the first term is obtained by averaging the Wishart density over draws on $\{\mathbf{b}_i\}$ from the full run. To estimate the second ordinate, which is conditioned on $\mathbf{D}^*$, we run a reduced MCMC simulation with the full conditional densities

$$\pi(\beta|\mathbf{y},\mathbf{D}^*,\sigma^2)\,;\,\pi(\sigma^2|\mathbf{y},\beta,\mathbf{D}^*,\{\mathbf{b}_i\})\,;\pi(\{\mathbf{b}_i\}|\mathbf{y},\beta,\mathbf{D}^*,\sigma^2) ,$$

where each conditional utilizes the fixed value of $\mathbf{D}$. The second ordinate is now estimated by averaging the inverse-gamma full conditional density of σ^2 at σ^{2*} over the draws on $(\beta,\{\mathbf{b}_i\})$ from this reduced run. The third ordinate is multivariate normal as given above and available directly.

15.9.2 Gaussian–Gaussian Tobit model

As another example, consider the Gaussian–Gaussian Tobit censored regression model. The likelihood ordinate is not available directly but can be estimated by a simulation-based approach. For the posterior ordinate we again utilize the decomposition

$$\pi(\mathbf{D}^{-1*},\sigma^{2*},\beta^*|\mathbf{y}) = \pi(\mathbf{D}^{-1*}|\mathbf{y})\pi(\sigma^{2*}|\mathbf{y},\mathbf{D}^*)\pi(\beta^*|\mathbf{y},\mathbf{D}^*,\sigma^{2*}) ,$$

where the first term is obtained by averaging the Wishart density over draws on $\{\mathbf{z}_i\}$ and $\{\mathbf{b}_i\}$ from the full run. To estimate the second ordinate, which is conditioned on $\mathbf{D}^*$, we run a reduced MCMC simulation with the full conditional densities

$$\begin{aligned}&\pi(\beta|\mathbf{y}_z,\mathbf{D}^*,\sigma^2)\,;\;\pi(\{\mathbf{z}_i\}|\mathbf{y},\beta,\mathbf{D}^*,\sigma^2);\\ &\pi(\sigma^2|\mathbf{y}_z,\beta,\mathbf{D}^*,\{\mathbf{b}_i\})\,;\,\pi(\{\mathbf{b}_i\}|\mathbf{y}_z,\beta,\mathbf{D}^*,\sigma^2) ,\end{aligned}$$

and estimate the second ordinate by averaging the inverse-gamma full conditional density of σ^2 at σ^{2*} over the draws on $(\beta,\{\mathbf{z}_i\},\{\mathbf{b}_i\})$ from this run. Finally, to

estimate the last ordinate we also fix σ^2 at σ^{2*} and continue the reduced runs with the full-conditional densities

$$\pi(\beta|\mathbf{y}_z, \mathbf{D}^*, \sigma^{2*});\ \pi(\{\mathbf{z}_i\}|\mathbf{y}, \beta, \mathbf{D}^*, \sigma^{2*});\ \pi(\{\mathbf{b}_i\}|\mathbf{y}_z, \beta, \mathbf{D}^*, \sigma^{2*}),$$

and average the multivariate normal density given in Step 1 of the MCMC algorithm at the point β^*.

15.9.3 Panel Poisson Model

As a last example of the calculation of the marginal likelihood, consider the panel poisson model in which the full conditional of β is not of known form. Now the posterior ordinate given the sampling scheme in the Panel count algorithm is decomposed as

$$\pi(\mathbf{D}^{-1*}, \beta^*|\mathbf{y}) = \pi(\mathbf{D}^{-1*}|\mathbf{y})\pi(\beta^*|\mathbf{y}, \mathbf{D}^*)$$

where the first ordinate is found by averaging the Wishart density over draws on $\{\mathbf{b}_i\}$ from the full run. The second ordinate is found by the method of Chib and Jeliazkov (2001) as

$$\hat{\pi}(\beta^*|\mathbf{y}, \mathbf{D}^*) = \frac{M^{-1}\sum_{g=1}^{M}\alpha(\beta^{(g)}, \beta^*|\mathbf{y}, \{\mathbf{b}_i^{(g)}\})q(\beta^*|\mathbf{y}, \{\mathbf{b}_i^{(g)}\})}{J^{-1}\sum_{j=1}^{J}\alpha(\beta^*, \beta^{(j)}|\mathbf{y}, \{\mathbf{b}_i^{(j)}\})}$$

where the draws in the numerator are from a reduced run comprising the full conditional distributions of β and $\{\mathbf{b}_i\}$, conditioned on $\mathbf{D}^*$ whereas the draws in the denominator are from a second reduced run comprising the full conditional distributions of $\{\mathbf{b}_i\}$, conditioned on $(\mathbf{D}^*, \beta^*)$ with an appended step in which $\beta^{(j)}$ is drawn from $q(\beta|\mathbf{y}, \{\mathbf{b}_i^{(j)}\})$. The log of the likelihood ordinate $p(\mathbf{y}|\beta^*, \mathbf{D}^*)$ is found by importance sampling.

15.10 Conclusion

In this chapter we have illustrated how Bayesian methods provide a complete inferential tool-kit for a variety of panel data models. The methods are based on a combination of hierarchical prior modeling and MCMC simulation methods. Interestingly, the approaches are able to tackle estimation and model comparison questions in situations that are quite challenging by other means. We discussed applications to models for continuous, binary, censored, count, multinomial response models under various realistic and robust distributional and modeling assumptions. The methods are quite practical and straightforward, even in complex models settings such as those with binary and count responses, and enable the calculation of the entire posterior distribution of the unknowns in the models. The algorithm for

fitting panel probit models with random effects is particularly interesting in that it highlights the value of augmentation in simplifying the simulations and in circumventing the calculation of the likelihood function. Procedures for dealing with missing data, predicting future outcomes and for detecting outliers have also been discussed.

The methods discussed in this chapter, which have arisen in the course of a revolutionary growth in Bayesian statistics in the last decade, offer a unified approach for analyzing a whole array of panel models. The pace of growth of Bayesian methods for longitudinal data continues unimpeded as the Bayesian approach attracts greater interest and adherents.

References

ALBERT, J. and S. CHIB (1993), Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, 88, 669–679.

ALBERT, J. and S. CHIB (1995), Bayesian residual analysis for binary response models, *Biometrika*, 82, 747–759.

BASU, S. and S. CHIB (2003), Marginal likelihood and Bayes factors for Dirichlet process mixture models, *Journal of the American Statistical Association*, 98, 224–235.

BESAG, J. (1974), Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society*, B, 36, 192–236.

BUTLER, J. S. and R. MOFFITT (1982), A computationally efficient quadrature procedure for the one factor multinomial probit model, *Econometrica*, 50, 761–764.

CARLIN, B. P. and T. LOUIS (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., Chapman and Hall, New York.

CHALONER, K. and R. BRANT (1988), A Bayesian approach to outlier detection and residual analysis, *Biometrika*, 75, 651–659.

CHAMBERLAIN, G. (1980), Analysis of covariance with qualitative data, *Review of Economic Studies*, 47, 225–238.

CHEN, Z. and D. B. DUNSON (2003), Random effects selection in linear mixed models, *Biometrics*, 59, 762–769.

CHEN, M-H and Q-M. SHAO (1998), On Monte Carlo methods for estimating ratios of normalizing constants, *Annals of Statistics*, 25, 1563–1594.

CHIANG, J., S. CHIB and C. NARASIMHAN (1999), Markov Chain Monte Carlo and models of consideration set and parameter heterogeneity, *Journal of Econometrics*, 89, 223–248.

CHIB, S. (1992), Bayes regression for the Tobit censored regression model, *Journal of Econometrics*, 51, 79–99.

CHIB, S. (1995), Marginal likelihood from the Gibbs output, *Journal of the American Statistical Association*, 90, 1313–1321.

CHIB, S. (2001), Markov Chain Monte Carlo methods: Computation and inference, in *Handbook of Econometrics* volume 5 (eds., J. J. Heckman and E. Leamer), Amsterdam: North Holland 3569-3649.

CHIB, S. and E. GREENBERG (1994), Bayes inference for regression models with ARMA(p,q) errors, *Journal of Econometrics*, 64, 183–206.

CHIB, S. and E. GREENBERG (1995), Understanding the Metropolis–Hastings algorithm, *American Statistician*, 49, 327–335.

CHIB, S. and E. GREENBERG (1996), Markov chain Monte Carlo simulation methods in econometrics, *Econometric Theory*, 12, 409–431.

CHIB, S. E. GREENBERG, and R. WINKLEMNANN (1998), Posterior simulation and Bayes factors in panel count data models, *Journal of Econometrics*, 86, 33–54.

CHIB, S. and B. P. CARLIN (1999), On MCMC sampling in hierarchical longitudinal models, *Statistics and Computing*, 9, 17–26.

CHIB, S. and I. JELIAZKOV (2001), Marginal likelihood from the Metropolis–Hastings output, *Journal of the American Statistical Association*, 96, 270–281.

CHIB, S., P. B. SEETHARAMAN, and A. STRIJNEV (2004) Model of brand choice with a no-purchase option calibrated to scanner panel data, *Journal of Marketing Research*, 184–196.

CONGDON, P. (2001) *Bayesian Statistical Modelling*, John Wiley & Sons, Chichester.

DICICCIO, T. J., R. E., KASS, A. E., RAFTERY and L. WASSERMAN (1997), Computing Bayes factors by combining simulation and asymptotic approximations, *Journal of the American Statistical Association*, 92, 903–915.

GELFAND, A. E. and A. F. M. SMITH (1990), Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, 85, 398–409.

GELFAND, A. E., S. K. SAHU, and B. P. CARLIN (1995), Efficient parameterizations for normal linear mixed models, *Biometrika*, 82, 479–488.

GEMAN, S. and D. GEMAN (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12, 609–628.

HASTINGS, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika,* 57, 97–109.

HECKMAN, J.J. (1981), Statistical models for discrete panel data, in *Structural Analysis of Discrete Data with Econometric Applications*, (eds., C. F. Manski and D. McFadden), Cambridge: MIT Press, 114–178.

LINDLEY, D.V. and A.F.M. SMITH (1972), Bayes estimates for the linear model, *Journal of the Royal Statistical Society* B, 34, 1–41.

METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, and E. TELLER (1953), Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, 21, 1087–1092.

PAGE, E (1977), Approximations to the cumulative normal function and its inverse for use on a pocket calculator, *Applied Statistics*, 26, 75–76.

ROBERTS, C.P. (2001), *The Bayesian Choice*, New York: Springer Verlag.

RIPLEY, B. (1987), *Stochastic Simulation*, New York: John Wiley & Sons.

STOUT, W. F. (1974), *Almost Sure Convergence*, New York: Academic Press.

TANNER, M. A. and W. H. WONG (1987), The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association,* 82, 528–549.

TIERNEY, L. (1994), Markov chains for exploring posterior distributions (with discussion), *Annals of Statistics*, 22, 1701–1762.

WAKEFIELD, J. C., A. F. M. SMITH, A. Racine POON, and A. E. GELFAND (1994), Bayesian analysis of linear and non-linear population models by using the Gibbs sampler, *Applied Statistics*, 43, 201–221.

ZELLNER, A (1975), Bayesian analysis of regression error terms, *Journal of the American Statistical Association*, 70, 138–144.

Chapter 16
To Pool or Not to Pool?

Badi H. Baltagi, Georges Bresson and Alain Pirotte

16.1 Introduction

For panel data studies with large N and small T, it is usual to pool the observations, assuming homogeneity of the slope coefficients. The latter is a testable assumption which is quite often rejected. Moreover, with the increasing time dimension of panel data sets, some researchers including Robertson and Symons (1992), Pesaran and Smith (1995), and Pesaran, Smith and Im (1996) have questioned the poolability of the data across heterogeneous units. Instead, they argue in favor of heterogeneous estimates that can be combined to obtain homogeneous estimates if the need arises. Maddala, Trost, Li and Joutz (1997) on the other hand argue that the heterogeneous time series estimates yield inaccurate estimates and even wrong signs for the coefficients, while the panel data estimates are not valid when one rejects the hypothesis of homogeneity of the coefficients. If one is after reliable coefficient estimates, Maddala, Trost, Li and Joutz (1997) argue in favor of shrinkage estimators that shrink the heterogeneous estimators towards the pooled homogeneous estimator. Proponents of the homogeneous panel estimators have acknowledged the potential heterogeneity among the cross-sectional units, but have assumed that the efficiency

Badi H. Baltagi
Center for Policy Research, 426 Eggers Hall, Syracuse University, Syracuse, NY 13244-1020, USA, e-mail: bbaltagi@maxwell.syr.edu

Georges Bresson
ERMES (UMR 7181, CNRS), Université Paris II and TEPP (FR 3126, CNRS), Institute for Labor Studies and Public Policies, 12, place du Panthéon, 75230 Paris Cedex 05, France, e-mail: bresson01@aol.com

Alain Pirotte
ERMES (UMR 7181, CNRS), Université Paris II and TEPP (FR 3126, CNRS), Institute for Labor Studies and Public Policies, 12, place du Panthéon, 75230 Paris Cedex 05, France, e-mail: apirotte@aol.com

* We dedicate this chapter to the memory of G.S. Maddala who wrote the original chapter with W. Hu.

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

gains from pooling outweighed these costs; see Hsiao (2003) on the benefits of panels. Clearly, in panel data sets with T very small, traditional homogeneous panel estimators would appear the only viable alternative. But as T reaches 50 years of post-war annual data, the choice no longer seems clear-cut.

In the context of dynamic demand for gasoline across 18 OECD countries over the period 1960–1990, Baltagi and Griffin (1997) argued for pooling the data as the best approach for obtaining reliable price and income elasticities. They also pointed out that pure cross-section studies cannot control for unobservable country effects, whereas pure time-series studies cannot control for unobservable oil shocks or behavioral changes occurring over time. Baltagi and Griffin (1997) compared the homogeneous and heterogeneous estimates in the context of gasoline demand based on the plausibility of the price and income elasticities as well as the speed of adjustment path to the long-run equilibrium. They found considerable variability in the parameter estimates among the heterogeneous estimators, some giving implausible estimates, while the homogeneous estimators gave similar plausible short-run estimates that differed only in estimating the long-run effects. Baltagi and Griffin (1997) also compared the forecast performance of these homogeneous and heterogeneous estimators over 1, 5, and 10 years horizon. Their findings show that the homogeneous estimators outperformed their heterogeneous counterparts based on mean squared forecast error. This result was replicated using a panel data set of 21 French regions over the period 1973–1998 by Baltagi, Bresson, Griffin and Pirotte (2003). Unlike the international OECD gasoline data set, the focus on the inter-regional differences in gasoline prices and income within France posed a different type of data set for the heterogeneity versus homogeneity debate. The variation in these prices and income were much smaller than international price and income differentials. This in turn reduces the efficiency gains from pooling and favors the heterogeneous estimators, especially given the differences between the Paris region and the rural areas of France. Baltagi, Bresson, Griffin and Pirotte (2003) showed that the time series estimates for each region are highly variable, unstable, and offer the worst out-of-sample forecasts. Despite the fact that the shrinkage estimators proposed by Maddala, Trost, Li and Joutz (1997) outperformed these individual heterogeneous estimates, they still had a wide range and were outperformed by the homogeneous estimators in out-of-sample forecasts. In addition, Baltagi, Griffin and Xiong (2000) carried out this comparison for a dynamic demand for cigarettes across 46 U.S. states over 30 years (1963–1992). Once again the homogeneous panel data estimators beat the heterogeneous and shrinkage type estimators in RMSE performance for out-of-sample forecasts. In another application, Driver, Imai, Temple and Urga (2004) utilize the Confederation of British Industry's (CBI) survey data to measure the impact of uncertainty on U.K. investment authorizations. The panel consists of 48 industries observed over 85 quarters 1978(Q1) to 1999(Q1). The uncertainty measure is based on the dispersion of beliefs across survey respondents about the general business situation in their industry. Following Baltagi and Griffin (1997) and Pesaran and Smith (1995), this paper questions the poolability of this data across different industries. The heterogeneous estimators considered are OLS and 2SLS at the industry level, as well as the unrestricted SUR estimation method. Fixed effects, random effects, pooled 2SLS, and restricted SUR are the homogeneous estimators considered.

The panel estimates find that uncertainty has a negative, non-negligible effect on investment, while the heterogeneous estimates vary considerably across industries. Forecast performance for 12 out-of-sample quarters 1996(Q2) to 1999(Q1) are compared. The pooled homogeneous estimators outperform their heterogeneous counterparts in terms of RMSE.

Maddala, Trost, Li and Joutz (1997) applied classical, empirical Bayes and Bayesian procedures to the problem of estimating short-run and long-run elasticities of residential demand for electricity and natural gas in the U.S. for 49 states over 21 years (1970–1990). Since the elasticity estimates for each state were the ultimate goal of their study they were faced with three alternatives. The first is to use individual time series regressions for each state. These gave bad results, were hard to interpret, and had several wrong signs. The second option was to pool the data and use panel data estimators. Although the pooled estimates gave the right signs and were more reasonable, Maddala, Trost, Li and Joutz (1997) argued that these estimates were not valid because the hypothesis of homogeneity of the coefficients was rejected. The third option, which they recommended, was to allow for some (but not complete) heterogeneity or (homogeneity). This approach lead them to their preferred shrinkage estimator which gave them more reasonable parameter estimates. In a follow-up study, Baltagi, Bresson and Pirotte (2002) reconsidered the two U.S. panel data sets on residential electricity and natural-gas demand used by Maddala, Trost, Li and Joutz (1997) and compared the out-of-sample forecast performance of the homogeneous, heterogeneous, and shrinkage estimators. Once again the results show that when the data is used to estimate heterogeneous models across states, individual estimates offer the worst out-of-sample forecasts. Despite the fact that shrinkage estimators outperform these individual estimates, they are outperformed by simple homogeneous panel data estimates in out-of-sample forecasts. Admittedly, there are additional case studies using U.S. data, but they do add to the evidence that simplicity and parsimony in model estimation offered by the homogeneous estimators yield better forecasts than the more parameter consuming heterogeneous estimators.

Proponents of the heterogeneous estimators include Pesaran and Smith (1995) and Pesaran, Shin and Smith (1999), who advocate abandoning the pooled approach altogether because of the inherent parameter heterogeneity, relying instead upon the average response from individual regressions. In fact, an earlier paper by Robertson and Symons (1992) studied the properties of some panel data estimators when the regression coefficients vary across individuals, i.e., they are heterogeneous but are assumed homogeneous in estimation. This was done for both stationary and nonstationary regressors. The basic conclusion was that severe biases can occur in dynamic estimation even for relatively small parameter variation. Using an empirical example of a real wage equation for a panel of 13 OECD countries observed over the period 1958–1986, Robertson and Symons (1992) show that parameter homogeneity across countries is rejected and the true relationship appears dynamic. Imposing false equality restriction biases the coefficient of the lagged wage upwards and the coefficient of the capital-labor ratio downwards.

Pesaran and Smith (1995) consider the problem of estimating a dynamic panel data model when the parameters are individually heterogeneous and illustrate their results by estimating industry-specific U.K. labor demand functions. In this case the model is given by

$$y_{it} = \lambda_i y_{i,t-1} + \beta_i x_{it} + u_{it} \quad i = 1, \ldots, N \quad t = 1, \ldots, T \tag{16.1}$$

where λ_i is i.i.d. $(\lambda, \sigma_\lambda^2)$ and β_i is i.i.d. (β, σ_β^2). Further λ_i and β_i are independent of y_{is}, x_{is}, and u_{is} for all s. The objective in this case is to obtain consistent estimates of the mean values of λ_i and β_i. Pesaran and Smith (1995) present four different estimation procedures:

(1) aggregate time-series regressions of group averages;
(2) cross-section regressions of averages over time;
(3) pooled regressions allowing for fixed or random intercepts;
(4) separate regressions for each group, where coefficients estimates are averaged over these groups.

They show that when T is small (even if N is large), all the procedures yield inconsistent estimators. The difficulty in obtaining consistent estimates for λ and β can be explained by rewriting the above equation as

$$y_{it} = \lambda y_{i,t-1} + \beta x_{it} + v_{it} \tag{16.2}$$

where $v_{it} = u_{it} + (\lambda_i - \lambda) y_{i,t-1} + (\beta_i - \beta) x_{it}$. By continuous substitution of $y_{i,t-s}$ it is easy to see that v_{it} is correlated with all present and past values of $y_{i,t-1-s}$ and $x_{i,t-s}$ for $s \geqslant 0$. The fact that v_{it} is correlated with the regressors renders the OLS estimator inconsistent, and the fact that v_{it} is correlated with $(y_{i,t-1-s}, x_{i,t-s})$ for $s > 0$ rules out the possibility of choosing any lagged value of y_{it} and x_{it} as legitimate instruments. When both N and T are large, Pesaran and Smith (1995) show that the cross-section regression procedure will yield consistent estimates of the mean values of λ and β. Intuitively, when T is large, the individual parameters λ_i and β_i can be consistently estimated using T observations of each individual i, say $\widehat{\lambda}_i$ and $\widehat{\beta}_i$; then, averaging these individual estimators, $\sum_{i=1}^{N} \widehat{\lambda}_i / N$ and $\sum_{i=1}^{N} \widehat{\beta}_i / N$, will lead to consistent estimators of the mean values of λ and β.

Hsiao and Tahmiscioglu (1997) use a panel of 561 U.S. firms over the period 1971–1992 to study the influence of financial constraints on company investment. They find substantial differences across firms in terms of their investment behavior. When a homogeneous pooled model is assumed, the impact of liquidity on firm investment is seriously underestimated. The authors recommend a mixed fixed and random coefficients framework based on the recursive predictive density criteria.

Pesaran, Smith and Im (1996) investigated the small sample properties of various estimators of the long-run coefficients for a dynamic heterogeneous panel data model using Monte Carlo experiments. Their findings indicate that the mean group estimator performs reasonably well for large T. However, when T is small, the mean group estimator could be seriously biased, particularly when N is large relative to T. Pesaran, Shin and Smith (1999) examine the effectiveness of alternative bias-correction procedures in reducing the small sample bias of these estimators using Monte Carlo experiments. An interesting finding is that when the coefficient of the lagged dependent variable is greater than or equal to 0.8, none of the bias correction procedures seem to work. Hsiao, Pesaran and Tahmiscioglu (1999) suggest a

Bayesian approach for estimating the mean parameters of a dynamic heterogeneous panel data model. The coefficients are assumed to be normally distributed across cross-sectional units and the Bayes estimator is implemented using Markov Chain Monte Carlo methods. Hsiao, Pesaran and Tahmiscioglu (1999) argue that Bayesian methods can be a viable alternative in the estimation of mean coefficients in dynamic panel data models even when the initial observations are treated as fixed constants. They establish the asymptotic equivalence of this Bayes estimator and the mean group estimator proposed by Pesaran and Smith (1995). The asymptotics are carried out for both N and $T \to \infty$ with $\sqrt{N}/T \to 0$. Monte Carlo experiments show that this Bayes estimator has better sampling properties than other estimators for both small and moderate size T. Hsiao, Pesaran and Tahmiscioglu (1999) also caution against the use of the mean group estimator unless T is sufficiently large relative to N. The bias in the mean coefficient of the lagged dependent variable appears to be serious when T is small and the true value of this coefficient is larger than 0.6. Hsiao, Pesaran and Tahmiscioglu (1999) apply their methods to estimate the q investment model using a panel of 273 U.S. firms over the period 1972–1993.

Depending on the extent of cross-sectional heterogeneity in the parameters, researchers may prefer these heterogeneous estimators to the traditional pooled homogeneous parameter estimators. In fact, Hsiao, Pesaran and Tahmiscioglu (1999) argued that there is not clarity in the literature about the appropriate estimation technique for dynamic panel data models, especially when the time series is short. They suggested a hierarchical Bayes approach to the estimation of such models using Markov Chain Monte Carlo methods (via Gibbs sampling).

By now, it is well known that pooling in the presence of parameter heterogeneity can produce misleading results. So, it is important to know if the pooling assumption is justified. Section 16.2 describes tests for poolability and Stein-rule methods. This is illustrated for a Tobin q investment application based on Hsiao and Tahmiscioglu (1997). Section 16.3 presents several heterogeneous estimators based on the sampling approach, the averaging approach and the Bayesian approach. Section 16.4 revisits the comparison of the out of sample forecast performance of the homogeneous and heterogeneous estimators in the context of the Tobin q application.

16.2 Tests for Poolability, Pretesting and Stein-Rule Methods

16.2.1 Tests for Poolability

The question of whether to pool the data or not naturally arises with panel data. The restricted model is the pooled model:

$$y_{it} = \alpha + X_{it}\beta + u_{it},\ i = 1,\ldots,N,\ t = 1,\ldots,T \tag{16.3}$$

which utilizes a one-way error component model for the disturbances:

$$u_{it} = \mu_i + v_{it} \tag{16.4}$$

where μ_i denotes the unobservable individual specific effect and v_{it} denotes the remainder disturbance. In vector form, (16.3) can be written as:

$$y = \alpha e_{NT} + X\beta + u = Z\theta + u \tag{16.5}$$

where y is $(NT \times 1)$, X is $(NT \times (k-1))$, $Z = [e_{NT}, X]$, $\theta' = (\alpha, \beta')$, e_{NT} is a vector of ones of dimension NT, and u is $(NT \times 1)$. Equation (16.4) can be written as:

$$u = Z_\mu \mu + v \tag{16.6}$$

where

$$Z_\mu = I_N \otimes e_T,\ \mu' = (\mu_1, \ldots, \mu_N) \text{ and } v \sim \text{i.i.d.}\left(0, \sigma_v^2 I_{NT}\right).$$

I_N is an identity matrix of dimension N, e_T is a vector of ones of dimension T, and $\otimes$ denotes the Kronecker product. If the μ_i are assumed to be fixed parameters, the model is called the fixed effects model. If the μ_i are assumed to be random (i.e., $\mu_i \sim \text{i.i.d.}\left(0, \sigma_\mu^2\right)$), the model is called the random effects model.

This pooled model represents a behavioral equation with the same parameters across individuals and over time. The unrestricted model, however, is the same behavioral equation but with different parameters across individuals and/or time. The question of whether to pool or not to pool boils down to the question of whether the parameters vary across individuals and/or over time. In what follows, we study the tests of poolability of the data for the case of pooling across individuals keeping in mind that the other case of pooling over time can be obtained in a similar fashion.

For the unrestricted model, we have a regression equation for each individual given by:

$$y_i = Z_i \theta_i + u_i,\ i = 1, \ldots, N \tag{16.7}$$

where y_i is $(T \times 1)$, $Z_i = [e_T, X_i]$, X_i is $(T \times k-1)$, $\theta_i' = (\alpha_i, \beta_i')$, and u_i is $(T \times 1)$. So, θ_i is different for every individual equation. We want to test the hypothesis

$$H_0 : \theta_i = \theta, \forall i.$$

So, under H_0, we can write the restricted model as: $y = Z\theta + u$. The unrestricted model can also be written as:

$$y = Z^* \theta^* + u = \begin{pmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & Z_N \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix} \tag{16.8}$$

where $Z = Z^* I^*$ with $I^* = (e_N \otimes I_k)$, e_N is a vector of ones of dimension N, and I_k is an identity matrix of dimension k.

16.2.1.1 Test for Poolability Under $u \sim N\left(0, \sigma_u^2 I_{NT}\right)$

We suppose that the disturbance u follows an normal distribution of zero mean and constant variance $\sigma_u^2 I_{NT}$, $\left(\text{i.e., } \mu_i = 0, \sigma_\mu^2 = 0\right)$. There is no individual specific effect. Then, the minimum variance unbiased (MVU) estimator for θ is the OLS estimator:

$$\widehat{\theta}_{\text{OLS}} = \widehat{\theta}_{\text{MLE}} = \left(Z'Z\right)^{-1} Z'y$$

and therefore

$$y = Z\widehat{\theta}_{\text{OLS}} + \widehat{u}_{\text{OLS}}.$$

Similarly, the MVU for θ_i is given by:

$$\widehat{\theta}_{i,\text{OLS}} = \widehat{\theta}_{i,\text{MLE}} = \left(Z_i'Z_i\right)^{-1} Z_i'y_i$$

and therefore

$$y_i = Z_i\widehat{\theta}_{i,\text{OLS}} + \widehat{u}_{i,\text{OLS}}.$$

Under H_0, the following test statistic:

$$F_{\text{obs}} = \frac{\left(\widehat{u}'_{\text{OLS}}\widehat{u}_{\text{OLS}} - \sum_{i=1}^{N} \widehat{u}'_{i,\text{OLS}}\widehat{u}_{i,\text{OLS}}\right) / (N-1)k}{\left(\sum_{i=1}^{N} \widehat{u}'_{i,\text{OLS}}\widehat{u}_{i,\text{OLS}}\right) / N(T-k)} \tag{16.9}$$

is distributed as $F\left((N-1)k, N(T-k)\right)$. Hence, the critical region for this test is defined as:

$$\left\{F_{\text{obs}} > F\left((N-1)k, N(T-k); \alpha_0\right)\right\}$$

where α_0 denotes the significance level of the test. This is exactly the Chow test extended to the case of N linear regressions. Therefore, if an economist has reason to believe that assumption $u \sim N\left(0, \sigma_u^2 I_{NT}\right)$ is true, and wants to pool his data across individuals, then it is recommended that he test for the poolability of the data using the Chow test given in (16.9). For an extension of the Chow test for poolability to a non-parametric panel data model that is robust to functional form misspecification, see Baltagi, Hidalgo, and Li (1996).

The problem with the Chow test is that $\Omega \neq \sigma_u^2 I_{NT}$. In fact, for the one-way error component model

$$\begin{aligned} \Omega &= E\left[uu'\right] = Z_\mu E\left[\mu\mu'\right] Z'_\mu + E\left[vv'\right] \\ &= \sigma_\mu^2 \left(I_N \otimes J_T\right) + \sigma_v^2 \left(I_N \otimes I_T\right) \end{aligned} \tag{16.10}$$

where $J_T = e_T e'_T$. Therefore, even if we assume normality on the disturbances, the Chow statistic will not have the F-distribution described above. However, a generalized Chow test which takes into account the general form of Ω will be the right test to perform. This is taken up in the next section.

16.2.1.2 Test for Poolability Under the General Assumption $u \sim N(0, \Omega)$

All we need to do is transform our model (under both the null and alternative hypothesis) such that the transformed disturbances have a variance of $\sigma_u^2 I_{NT}$, then apply the Chow test on the transformed model. Given $\Omega = \sigma_u^2 \Sigma$, we premultiply the restricted model by $\Sigma^{-1/2}$ and call $\Sigma^{-1/2} y = \widetilde{y}$. Hence:

$$\widetilde{y} = \widetilde{Z}\theta + \widetilde{u}$$

with $E[\widetilde{u}\widetilde{u}'] = \Sigma^{-1/2} E[uu'] \Sigma^{-1/2'} = \sigma_u^2 I_{NT}$. Similarly, we premultiply the unrestricted model (16.8) by $\Sigma^{-1/2}$ and call $\Sigma^{-1/2} Z^* = \widetilde{Z}^*$. Therefore,

$$\widetilde{y} = \widetilde{Z}^* \theta^* + \widetilde{u}$$

with $E[\widetilde{u}\widetilde{u}'] = \sigma_u^2 I_{NT}$. At this stage, we can test $H_0 : \theta_i = \theta$ for every $i = 1, 2, \ldots, N$, simply by using the Chow test only now on the transformed models. Under H_0, the following test statistic:

$$\widetilde{F}_{\text{obs}} = \frac{y' \left[\Sigma^{-1} \left(Z^* \left(Z^{*\prime} \Sigma^{-1} Z^* \right)^{-1} Z^{*\prime} - Z \left(Z' \Sigma^{-1} Z \right)^{-1} Z' \right) \Sigma^{-1} \right] y / (N-1)k}{\left(y' \Sigma^{-1} y - y' \Sigma^{-1} Z^* \left(Z^{*\prime} \Sigma^{-1} Z^* \right)^{-1} Z^{*\prime} \Sigma^{-1} y \right) / N(T-k)} \tag{16.11}$$

is distributed as $F((N-1)k, N(T-k))$. It is important to emphasize that (16.11) is operational only when Σ is known. This test is a special application of a general test for linear restrictions described by Roy in 1957 and used by Zellner in 1962 to test for aggregation bias in a set of seemingly unrelated regressions. In case Σ is unknown, we replace Σ in (16.11) by a consistent estimator (say $\widehat{\Sigma}$) and call the resulting test statistic $\widehat{\widetilde{F}}_{\text{obs}}$.

One of the main motivations behind pooling a time series of cross-sections is to widen our database in order to get better and more reliable estimates of the parameters of our model. Using the Chow test, the question of whether "to pool or not to pool" is reduced to a test of the validity of the null hypothesis $H_0 : \theta_i = \theta$ for all i. Imposing these restrictions (true or false) will reduce the variance of the pooled estimator, but may introduce bias if these restrictions are false. Baltagi (2005, pp. 54–58) discusses three mean squared error criteria suggested by Wallace (1972) for $\Omega = \sigma_u^2 I_{NT}$ and by McElroy (1977) for $\Omega = \sigma_u^2 \Sigma$. These MSE criteria do not test H_0, but rather help us choose on pragmatic grounds between the restricted pooled model and the unrestricted heterogeneous model. Using Monte Carlo experiments, Baltagi (1981) shows that the Chow test performs poorly, rejecting poolability when true under a random error component model whenever the variance components are large. Weaker MSE criteria reduced the frequency of type I error committed by the Chow test. However, the weaker MSE criteria performance was still poor compared to the Roy–Zellner test or the extensions of these weaker MSE criteria for a general Ω.

Recently, Bun (2004) focused on testing the poolability hypothesis across cross-section units assuming constant coefficients over time. In particular, this testing applies to panel data with a limited number of cross-section units, like countries or states observed over a long time period, i.e., with T larger than N. Bun (2004) uses Monte Carlo experiments to examine the actual size of various asymptotic procedures for testing the poolability hypothesis. Dynamic regression models as well as nonspherical disturbances are considered. Results show that the classical asymptotic tests have poor finite sample performance, while their bootstrapped counterparts lead to more accurate inference. An empirical example is given using panel data on GDP growth and unemployment rates in 14 OECD countries over the period 1966–1990. For this data set, it is shown that the classical asymptotic tests reject poolability while their bootstrap counterparts do not.

16.2.2 Pretesting and Stein-Rule Methods

Choosing a pooled estimator if we do not reject $H_0: \theta_i = \theta$ for all i, and the heterogeneous estimator if we reject H_0 leads to a pretest estimator.[1] This brings into question the appropriate level of significance to use with this preliminary test. In fact, the practice is to use significance levels much higher than 5%; see Maddala and Hu (1996).

Another problem with the pretesting procedure is that its sampling distribution is complicated; see Judge and Bock (1978). Also, these pretest estimators are dominated by Stein-rule estimators under quadratic loss function. Using a wilderness recreation demand model, Ziemer and Wetzstein (1983) show that a Stein-rule estimator gives better forecast risk performance than the pooled ($\widehat{\theta}_{\text{OLS}}$) or individual estimators ($\widehat{\theta}_{i,\text{OLS}}$). The Stein-rule estimator is given by:

$$\widehat{\theta}_i^S = \left(\frac{c}{F_{\text{obs}}}\right)\widehat{\theta}_{\text{OLS}} + \left(1 - \frac{c}{F_{\text{obs}}}\right)\widehat{\theta}_{i,\text{OLS}}. \tag{16.12}$$

The optimal value of the constant c suggested by Judge and Bock (1978) is:

$$c = \frac{(N-1)k-2}{N(T-k)+2}.$$

Note that $\widehat{\theta}_i^S$ shrinks $\widehat{\theta}_{i,\text{OLS}}$ towards the pooled estimator $\widehat{\theta}_{\text{OLS}}$ explaining why this estimator is often called Stein-rule shrinkage estimator. When N is large, the factor c is roughly $k/(T-k)$. If, in addition, the number of explanatory variables k is small relative to the number of time periods T, c will be small and, for a given F_{obs}, the shrinkage factor towards the pooled estimator (c/F_{obs}) will be small. The

[1] Hsiao and Pesaran (2007) present several Hausman type tests for slope heterogeneity based on the mean group estimator proposed by Pesaran, Shin and Smith (1999) (see the Sect. 16.3.1).

Bayesian and empirical Bayesian methods (discussed later) imply shrinking towards a weighted mean of the $\widehat{\theta}_i$ and not the pooled estimator $\widehat{\theta}$.

16.2.3 Example

We illustrate the tests of poolability and the Stein-rule method using a simple dynamic version of the classical Tobin q investment model studied by Hsiao and Tahmiscioglu (1997):

$$\left(\frac{I}{K}\right)_{it} = \alpha_i + \beta_{1i}\left(\frac{I}{K}\right)_{it-1} + \beta_{2i} q_{it} + u_{it} \tag{16.13}$$

or in vector form

$$y_{it} = Z_{it}\theta_i + u_{it}$$

where I_{it} denotes investment expenditures by firm i during period t, K_{it} is the replacement value of the capital stock, and q_{it} is Tobin's q of the firm. Tobin's q theory relates investment to marginal q, which is the ratio of the market value of new investment goods to their replacement cost. Thus, investment will be an increasing function of marginal q. Because marginal q is unobservable, it is common in empirical work to substitute it with average or Tobin's q. If a firm has unexploited profit opportunities, then an increase of its capital stock price of 1 unit will increase its market value by more than one unit ($q > 1$). Firms can be expected to increase investment until marginal q equals 1. On the other hand, if a firm has already more than adequate capital, then an increase in capital stock by one unit will increase its market value by less than one unit ($q < 1$). β_{1i} is the investment inertia of firm i and $(1-\beta_{1i})$ is the speed of adjustment. The panel data set used in this study contains 337 U.S. firms over 17 years (1982–1998).[2] Hsiao, Pesaran and Tahmiscioglu (1999) argued that the troubling findings of large estimates often obtained for the adjustment cost parameters and the implied slow speeds of adjustment of the capital stock to its equilibrium value may be due, at least partly, to the use of inappropriate estimation techniques when there is significant heterogeneity in firm investment responses to the q variable. The restricted model is:

$$\left(\frac{I}{K}\right)_{it} = \alpha + \beta_1\left(\frac{I}{K}\right)_{it-1} + \beta_2 q_{it} + u_{it} \tag{16.14}$$

or in vector form

$$y_{it} = Z_{it}\theta + u_{it}.$$

[2] This was kindly provided by Cheng Hsiao and A. Kamil Tashmiscioglu. This is not exactly the same data set as the one used by Hsiao, Pesaran and Tahmiscioglu (1999) which contains 273 firms over the 20 years (1973–1992). For a detailed description of these variables, see Hsiao and Tahmiscioglu (1997).

Table 16.1 Tests of poolability for the Tobin q investment model[3] $N = 337$, $T = 16$

	OLS	LSDV	FGLS
α	0.0872 (25.695)	—	0.0926 (2.585)
β_1	0.4469 (39.138)	0.2777 (22.136)	0.3252 (26.929)
β_2	0.0079 (18.829)	0.0157 (23.180)	0.0123 (21.838)
Chow test (intercept and slopes)		$2.6026 \sim F(1008, 4381)$	
Chow test (slopes only)		$2.4234 \sim F(672, 4718)$	
Roy–Zellner test		$1.5796 \sim F(1008, 4381)$	
Stein-rule shrinkage factor		0.9118	

We want first to check whether coefficients are constant or not across firms, that is if $H_0 : \theta_i = \theta$ for all i.

Table 16.1 shows that the Chow test for poolability across firms gives an observed F-statistic of 2.6026 which is distributed as $F(1008, 4718)$ under $H_0 : \theta_i = \theta$ for $i = 1, \ldots, N$. There are 1008 restrictions and the test rejects poolability across firms for all the coefficients. One can test for poolability of slopes only, allowing for varying intercepts. The restricted model is the within regression with firm dummies (LSDV). The observed F-statistic of 2.4234 is distributed as $F(672, 4718)$ under $H_0 : \beta_{ji} = \beta_j$ for $j = 1, 2$ and $i = 1, \ldots, N$. This again is significant at the 5% level and rejects the poolability of the slopes across firms. The Roy–Zellner test for poolability across firms, allowing for one-way error component disturbances, yields an observed F-value of 1.5796 and is distributed as $F(1008, 4381)$ under $H_0 : \theta_i = \theta$ for $i = 1, \ldots, N$. This still rejects poolability across firms even after allowing for one-way error component disturbances. The Stein-rule shrinkage factor is 91.18%, so the Stein-rule estimator $\widehat{\theta}_i^S$ is a linear combination of 8.82% weight on the pooled estimator $\widehat{\theta}_{\text{OLS}}$ and 91.18% on the heterogeneous estimator $\widehat{\theta}_{i,\text{OLS}}$.

16.3 Heterogeneous Estimators

When the data do not support the hypothesis of fixed coefficients, it would seem reasonable to allow for variations in the parameters across cross-sectional units. For the ith individual, a single-equation model can be written as:

$$y_i = Z_i \theta_i + u_i, \; i = 1, \ldots, N \tag{16.15}$$

that is:

$$y_i \sim N(Z_i \theta_i, \Omega_i) \tag{16.16}$$

[3] t-Statistics are in parentheses.

with $\Omega_i = E[u_i u_i'] = \sigma_i^2 I_T$. If all the parameters are treated as fixed and different for cross-sectional units, there will be Nk parameters with only NT observations. Obviously, there is no way we can obtain any meaningful estimates of θ_i, especially when k is closed to T. Alternatively, each regression coefficient can be viewed as a random variable with a probability distribution. The random coefficients specification reduces the number of parameters to be estimated substantially, while still allowing the coefficients to differ from unit to unit and/or from time to time. θ_i are assumed to be independently normally distributed with mean $\overline{\theta}$ and covariance Δ (with $\text{Cov}(\theta_i, \theta_j) = 0$, $i \neq j$):

$$\theta_i = \overline{\theta} + \varepsilon_i,\ \theta_i \sim N\left(\overline{\theta}, \Delta\right). \tag{16.17}$$

Substituting $\theta_i = \overline{\theta} + \varepsilon_i$ into (16.15) yields:

$$y_i = Z_i \overline{\theta} + v_i \tag{16.18}$$

where $v_i = Z_i \varepsilon_i + u_i$. This leads us to the Swamy (1970) model.

Stacking all NT observations, we have:

$$y = Z\overline{\theta} + v \tag{16.19}$$

where $v = Z^* \varepsilon + u$. The covariance matrix for the composite disturbance term v is bloc-diagonal, and is defined by

$$V[v] = \Sigma = \text{diag}\left(\Sigma_1, \ldots, \Sigma_N\right)$$

where

$$\Sigma_i = Z_i \Delta Z_i^{'} + \sigma_i^2 I_T.$$

The best linear unbiased estimator of $\overline{\theta}$ for (16.19) is the GLS estimator:

$$\widehat{\overline{\theta}}_{\text{GLS}} = \sum_{i=1}^{N} \Theta_i \widehat{\theta}_{i,\text{OLS}} \tag{16.20}$$

where

$$\widehat{\theta}_{i,\text{OLS}} = \left(Z_i' Z_i\right)^{-1} Z_i' y_i \tag{16.21}$$

and

$$\Theta_i = \left(\sum_{i=1}^{N} \left[\Delta + \sigma_i^2 \left(Z_i' Z_i\right)^{-1}\right]^{-1}\right)^{-1} \left[\Delta + \sigma_i^2 \left(Z_i' Z_i\right)^{-1}\right]^{-1}. \tag{16.22}$$

The covariance matrix for the GLS estimator is:

$$V\left[\widehat{\overline{\theta}}_{\text{GLS}}\right] = \left(\sum_{i=1}^{N} \left[\Delta + \sigma_i^2 \left(Z_i' Z_i\right)^{-1}\right]^{-1}\right)^{-1}. \tag{16.23}$$

Swamy proposed to use the least-squares estimators $\widehat{\theta}_i$ and their residuals $\widehat{r}_i \left(= y_i - Z_i \widehat{\theta}_{i,\mathrm{OLS}}\right)$ to obtain unbiased estimators of σ_i^2 and Δ (see Hsiao and Pesaran (2007)). Swamy (1970) provides an asymptotic normal and efficient estimator of the mean coefficients. Pesaran and Smith (1995), and Pesaran, Shin and Smith (1999) advocate alternative estimators which they call respectively the Mean Group estimator and the Pooled Mean Group estimator.

16.3.1 Averaging Estimators

The Mean Group estimator is obtained by estimating the coefficients of each cross-section separately by OLS and then taking an arithmetic average:

$$\widetilde{\overline{\theta}} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\theta}_{i,\mathrm{OLS}}. \tag{16.24}$$

When $T \to \infty$, $\widehat{\theta}_{i,\mathrm{OLS}} \to \theta_i$ and (16.24) will be consistent when N also goes to infinity. This estimator has obviously only asymptotic justification. However, it would be interesting to have some idea about its performance in finite sample, particularly as compared to Bayesian type estimators.

Pesaran, Shin and Smith (1999) proposed an estimator called the Pooled Mean Group estimator which constrains the long-run coefficients to be the same among individuals. Suppose that we want to estimate an ADL model:

$$y_{it} = Z_{it}\theta_i + u_{it} = \sum_{j=1}^{p} \lambda_{ij} y_{i,t-j} + \sum_{j=0}^{q} \delta_{ij}' x_{i,t-j} + \gamma_i' d_t + u_{it} \tag{16.25}$$

where x_{it} is a $(k \times 1)$ vector of explanatory variables and d_t is a $(s \times 1)$ vector of observations on fixed regressors such as intercept and time trends or variables that vary only over time. We can re-parametrize (16.25):

$$\Delta y_{it} = \phi_i y_{i,t-1} + \beta_i' x_{it} + \sum_{j=1}^{p-1} \lambda_{ij}^* \Delta y_{i,t-j} + \sum_{j=0}^{q-1} \delta_{ij}^{*\prime} \Delta x_{i,t-j} + \gamma_i' d_t + u_{it} \tag{16.26}$$

where

$$\phi_i = -\left(1 - \sum_{j=1}^{p} \lambda_{ij}\right),\ \beta_i = \sum_{j=0}^{q} \delta_{ij},\ \lambda_{ij}^* = -\sum_{m=j+1}^{p} \lambda_{im} \text{ and } \delta_{ij}^* = -\sum_{m=j+1}^{q} \delta_{im}.$$

If we stack the time series observations for each group, the error-correction model (16.26) becomes:

$$\underset{(T,1)}{\Delta y_i} = \phi_i \underset{(T,1)}{y_{i,-1}} + \beta_i' \underset{(T,k)}{X_i} + \sum_{j=1}^{p-1} \lambda_{ij}^* \underset{(T,1)}{\Delta y_{i,-j}} + \sum_{j=0}^{q-1} \underset{(T,k)}{\delta_{ij}^{*\prime} \Delta X_{i,-j}} + \gamma_i' D_t + \varepsilon_i \tag{16.27}$$

where $y_i = (y_{i1}, \ldots, y_{iT})'$, $X_i = (x_{i1}, \ldots, x_{iT})'$, and $D_t = (d_1, \ldots, d_T)'$. If $\phi_i < 0$, there exists a long-run relationship between y_{it} and x_{it} defined by:

$$y_{it} = \left(-\frac{\beta_i'}{\phi_i} \right) x_{it} + v_{it} \ , \forall i. \tag{16.28}$$

Pesaran, Shin and Smith (1999) constrain the long-run coefficients on X_i defined by $\omega_i = (-\beta_i/\phi_i)$ to be the same across individuals or across groups of individuals:

$$\omega_i = \omega \ , \forall i. \tag{16.29}$$

So, the ECM can be written more compactly as:

$$\Delta y_i = \phi_i \xi_i(\theta) + W_i \kappa_i + u_i \tag{16.30}$$

where the error correction component is:

$$\xi_i(\theta) = y_{i,-1} - X_i \omega$$

and

$$W_i = \left(\Delta y_{i,-1}, \ldots, \Delta y_{i,-p+1}, \Delta X_i, \Delta X_{i,-1}, \ldots, \Delta X_{i,-q+1}, D \right)$$
$$\kappa_i = \left(\lambda_{i1}^*, \lambda_{i2}^*, \ldots, \lambda_{ip-1}^*, \delta_{i0}^{*\prime}, \delta_{i1}^{*\prime}, \ldots, \delta_{iq-1}^{*\prime}, \gamma_i' \right)'.$$

If the disturbances are normally distributed, the ML estimation of the long-run coefficients ω and the individual-specific error-correction coefficients ϕ_i are obtained by maximizing the concentrated likelihood (see Pesaran, Shin and Smith (1999)).

16.3.2 Bayesian Framework

The underlying probability interpretation for a Bayesian is a subjective one, referring to a personal degree of belief. The rules of probability calculus are used to examine how prior beliefs are transformed to posterior beliefs by incorporating data information.[4] Here we only consider cases where the model parameter vector θ is of finite dimension. A Bayesian then focuses on the inference of θ (treated as a random variable) conditional on y and the underlying model M, summarized in the posterior density $p\left(\overline{\theta}|y, M\right)$. The observations in y define a mapping from the prior

[4] See Chib (2001).

$p\left(\overline{\theta}\right)$ into $p\left(\overline{\theta}|y,M\right)$. The posterior distribution of $\overline{\theta}$ can be derived by expressing the likelihood function conditional on the initial values y_{i0} and combining it with the prior distribution of $\overline{\theta}$:

$$p\left(\overline{\theta}|y,y_{i0}\right) \propto p\left(y|\overline{\theta}\right) p\left(\overline{\theta}\right).$$

Lindley and Smith (1972) proposed a three-stage hierarchy. The joint density function of the data y is such that:

$$y \sim N\left(Z^{*}\theta,\Omega\right) \text{ where } \Omega = E\left[uu'\right]$$

is indexed by a k-vector θ of unknown parameters assumed to be normally distributed

$$\theta \sim N\left(\overline{\theta},\Delta\right).$$

The third stage of the hierarchy corresponds to the prior distribution of $\overline{\theta}$

$$\overline{\theta} \sim N\left(\varphi,\Psi\right).$$

Using the properties of the multivariate normal distribution, we can define the conditional density of y given $\overline{\theta}$. If $y \sim N\left(Z^{*}\theta,\Omega\right)$ and $\theta \sim N\left(\overline{\theta},\Delta\right)$, then the marginal distribution of y conditional on $\overline{\theta}$ is

$$y \sim N\left(Z^{*}\overline{\theta},\Sigma\right)$$

where $\Sigma = \Omega + Z^{*}\Delta\, Z^{*\prime}$. Combining this with the prior distribution of $\overline{\theta}$ yields to the posterior density of $\overline{\theta}$. Then, the posterior density is proportional to:

$$p\left(\overline{\theta}|y,y_{i0}\right) \propto \exp\left(-\frac{1}{2}\left(y-Z^{*}\overline{\theta}\right)'\Sigma^{-1}\left(y-Z^{*}\overline{\theta}\right)-\frac{1}{2}\left(\overline{\theta}-\varphi\right)'\Psi^{-1}\left(\overline{\theta}-\varphi\right)\right).$$

Assuming prior ignorance at the third stage of the hierarchy (i.e., $\Psi^{-1}=0$) yields to the following posterior distribution of $\overline{\theta}$:

$$\overline{\theta} \sim N\left(\left(Z^{*\prime}\Sigma^{-1}Z^{*}\right)^{-1}Z^{*\prime}\Sigma^{-1}y,\left(Z^{*\prime}\Sigma^{-1}Z^{*}\right)^{-1}\right)$$

From a frequentist point of view and in order to estimate $\theta,\overline{\theta},\sigma_i^2$ and Δ, we must theoretically maximize the following log likelihood (see Maddala, Trost, Li and Joutz (1997)):

$$\begin{aligned}\operatorname{Log} L\left(\theta, \overline{\theta}, \sigma_i^2, \Delta \mid y, Z^*\right) = Cst &- \frac{T}{2}\sum_{i=1}^{N}\operatorname{Log}\sigma_i^2 - \frac{N}{2}\operatorname{Log}|\Delta| \\ &- \frac{1}{2}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(y_i - Z_i\theta_i\right)'\left(y_i - Z_i\theta_i\right) \\ &- \frac{1}{2}\sum_{i=1}^{N}\left(\theta_i - \overline{\theta}\right)'\Delta^{-1}\left(\theta_i - \overline{\theta}\right)\end{aligned}$$

we get

$$\widehat{\sigma}_i^2 = \frac{1}{T}\left(y_i - Z_i\theta_i\right)'\left(y_i - Z_i\theta_i\right)$$

$$\widehat{\Delta} = \frac{1}{N}\sum_{i=1}^{N}\left(\theta_i - \overline{\theta}\right)\left(\theta_i - \overline{\theta}\right)'$$

$$\widehat{\overline{\theta}} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}_i$$

and

$$\widehat{\theta}_i = \left[\widehat{\Delta}^{-1} + \frac{1}{\widehat{\sigma}_i^2}Z_i'Z_i\right]^{-1}\left[\widehat{\Delta}^{-1}\widehat{\overline{\theta}} + \frac{1}{\widehat{\sigma}_i^2}Z_i'Z_i\widehat{\theta}_{i,\mathrm{OLS}}\right].$$

which is the same as

$$\widehat{\theta}_i = \widehat{\overline{\theta}} + \widehat{\Delta}\,Z_i'\left[Z_i\widehat{\Delta}\,Z_i' + \widehat{\sigma}_i^2 I_T\right]^{-1}\left[y_i - Z_i\widehat{\overline{\theta}}\right]$$

For estimating Maximum Likelihood parameters $\widehat{\theta}_i$, $\widehat{\Delta}$ and $\widehat{\sigma}_i^2$, we must run a first step and use the OLS estimator for each individual.

The traditional approach to estimating regression coefficients with panel data is a dichotomy of either estimating θ_i from the data on the ith cross-section unit or from the pooled sample. The general solution that emerges from the Bayesian approach is to shrink each individual θ_i from the ith cross-section towards a common estimate $\overline{\theta}$.

We suppose that

$$\theta_i \sim N\left(\overline{\theta}, \Delta\right).$$

This statement defines the prior distribution of θ_i. The parameters $\overline{\theta}$ and Δ are unknown, then we must make some assumptions. After this, we can obtain the posterior distribution of θ_i. If $\overline{\theta}$, σ_i^2 and Δ were known, then the posterior distribution of θ_i is given by:

$$\widehat{\theta}_i^* = \left[\Delta^{-1} + \frac{1}{\sigma_i^2}Z_i'Z_i\right]^{-1}\left[\Delta^{-1}\overline{\theta} + \frac{1}{\sigma_i^2}Z_i'Z_i\widehat{\theta}_{i,\mathrm{OLS}}\right]$$

and its variance:

$$V\left[\widehat{\theta}_i^*\right] = \left[\Delta^{-1} + \frac{1}{\sigma_i^2} Z_i' Z_i\right]^{-1}$$

where $\widehat{\theta}_{i,\mathrm{OLS}}$ is the OLS estimator of θ_i. The posterior distribution mean of $\overline{\theta}$ is defined by:

$$\widehat{\overline{\theta}}^* = \frac{1}{N}\sum_{i=1}^{N} \widehat{\theta}_i^*.$$

Lindley and Smith (1972) have shown that prior distributions for nuisance parameters (including the variance–covariance matrix of the hyperparameters like $\overline{\theta}$) lead to integrals which cannot be all expressed in closed form. They suggest an approximation which consists of using the mode of the posterior distribution rather than the mean.[5] The former empirical Bayes estimator has been followed by other empirical Bayes methods such as iterative Bayes and empirical iterative Bayes estimators (see Maddala, Trost, Li and Joutz (1997) and Table 16.2).

A more flexible tool is the rejection sampling method discussed in Gelfand and Smith (1990, 1992) when the only requirement is that the maximum value of sampling density $\underset{\theta}{\mathrm{Max}}\ p(y|\theta, M)$ be finite. Hence, if we can draw from the prior $p(\theta)$, we can generate drawings from the posterior $p(\theta|y, M)$ simply by rejection. Markov Chain Monte Carlo (MCMC) simulations versions of this accept–reject algorithm have been recently proposed. As underlined by Chib (2001, 2007), these methods have revolutionized Bayesian statistics. One very popular MCMC method, introduced by Gelfand and Smith (1990), is called the Gibbs sampling method. Therefore, a full Bayesian implementation of the model is now feasible using sampling-based approaches to calculate marginal densities. Using Gibbs sampling, Hsiao, Pesaran and Tahmiscioglu (1999) have proposed the "Hierarchical Bayes" estimator.

16.3.2.1 Iterative Bayes Estimator

In general, σ_i^2 are Δ unknown parameters. Then we must make some prior assumptions about these parameters. Smith (1973) proposed for Δ^{-1} a conjugate Wishart distribution and for σ_i^2 some independent inverse χ^2 distributions (see the Sect. 16.3.2.4). Instead of the mean of the distribution, he used the posterior mode of the distribution:

$$\widehat{\sigma}_i^2 = \frac{1}{T + \varsigma_i + 2}\left[\varsigma_i \lambda_i + \left(y_i - Z_i \widehat{\theta}_i^*\right)' \left(y_i - Z_i \widehat{\theta}_i^*\right)\right] \tag{16.31}$$

and

$$\widehat{\Delta}^* = \frac{1}{N - k - 2 + \delta}\left[R + \sum_{i=1}^{N}\left(\widehat{\theta}_i^* - \widehat{\overline{\theta}}^*\right)\left(\widehat{\theta}_i^* - \widehat{\overline{\theta}}^*\right)'\right] \tag{16.32}$$

[5] This approximation is likely to be good only if the samples are fairly large and the resulting posterior distributions approximatively normal.

where ς_i, λ_i , δ, and R are hyperparameters of the prior distribution. Smith (1973) proposes to approximate these hyperparameters by using $\varsigma_i = 0$, $\delta = 1$ and R is a diagonal matrix with small positive entries $(= 0.001)$. The estimators are:

$$\widehat{\sigma}_i^2 = \frac{1}{T+2}\left(y_i - Z_i\widehat{\theta}_i^*\right)'\left(y_i - Z_i\widehat{\theta}_i^*\right) \tag{16.33}$$

$$\widehat{\Delta}^* = \frac{1}{N-k-1}\left[R + \sum_{i=1}^{N}\left(\widehat{\theta}_i^* - \widehat{\overline{\theta}}^*\right)\left(\widehat{\theta}_i^* - \widehat{\overline{\theta}}^*\right)'\right] \tag{16.34}$$

$$\widehat{\theta}_i^* = \left[\widehat{\Delta}^{*-1} + \frac{1}{\widehat{\sigma}_i^2}Z_i'Z_i\right]^{-1}\left[\widehat{\Delta}^{*-1}\widehat{\overline{\theta}}^* + \frac{1}{\widehat{\sigma}_i^2}Z_i'Z_i\widehat{\theta}_{i,\mathrm{OLS}}\right] \tag{16.35}$$

and

$$\widehat{\overline{\theta}}^* = \frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}_i^*. \tag{16.36}$$

The equations (16.33) to (16.36) must be estimated by iterative procedure. The initial iteration use OLS estimates.

16.3.2.2 Empirical Bayes Estimator

This estimator has been proposed by Smith (1973). It is a quite different as empirical Bayes's estimator proposed by Rao (1975). It is defined as:

$$\widehat{\overline{\theta}}^* = \frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}_{i,\mathrm{OLS}}$$

$$\widehat{\sigma}_i^2 = \frac{1}{T-k}\left(y_i'y_i - y_i'Z_i\widehat{\theta}_{i,\mathrm{OLS}}\right)$$

$$\widehat{\Delta}^* = \frac{1}{N-1}\left[\sum_{i=1}^{N}\left(\widehat{\theta}_{i,\mathrm{OLS}} - \widehat{\overline{\theta}}^*\right)\left(\widehat{\theta}_{i,\mathrm{OLS}} - \widehat{\overline{\theta}}^*\right)'\right]$$

and

$$\widehat{\theta}_i^* = \left[\widehat{\Delta}^{*-1} + \frac{1}{\widehat{\sigma}_i^2}Z_i'Z_i\right]^{-1}\left[\widehat{\Delta}^{*-1}\widehat{\overline{\theta}}^* + \frac{1}{\widehat{\sigma}_i^2}Z_i'Z_i\widehat{\theta}_{i,\mathrm{OLS}}\right].$$

This estimator is based on OLS estimates. The estimators of $\widehat{\sigma}_i^2$ and $\widehat{\Delta}^*$ are unbiased if Z_i contains only exogenous variables.

16.3.2.3 Empirical Iterative Bayes Estimator

This estimator was proposed by Maddala, Trost, Li and Joutz (1997). The parameters $\widehat{\sigma}_i^2$ and $\widehat{\Delta}^*$ are estimated by:

$$\widehat{\sigma}_i^2 = \frac{1}{T-k}\left(y_i - Z_i\widehat{\theta}_i^*\right)'\left(y_i - Z_i\widehat{\theta}_i^*\right)$$

and

$$\widehat{\Delta}^* = \frac{1}{N-1}\left[R + \sum_{i=1}^{N}\left(\widehat{\theta}_i^* - \widehat{\overline{\theta}}^*\right)\left(\widehat{\theta}_i^* - \widehat{\overline{\theta}}^*\right)'\right];$$

then we can compute

$$\widehat{\theta}_i^* = \left[\widehat{\Delta}^{*-1} + \frac{1}{\widehat{\sigma}_i^2}Z_i'Z_i\right]^{-1}\left[\widehat{\Delta}^{*-1}\widehat{\overline{\theta}}^* + \frac{1}{\widehat{\sigma}_i^2}Z_i'Z_i\widehat{\theta}_{i,\mathrm{OLS}}\right]$$

and

$$V\left[\widehat{\theta}_i^*\right] = \left[\widehat{\Delta}^{*-1} + \frac{1}{\widehat{\sigma}_i^2}Z_i'Z_i\right]^{-1}.$$

For the first iteration, we use the OLS estimates. Maddala, Trost, Li and Joutz (1997) argue that the iterative process for estimating Δ and $\overline{\theta}$ will yield to more efficient estimates of these parameters.

16.3.2.4 Hierarchical Bayes Estimator

We have previously seen that Lindley and Smith (1972) have proposed a three-stage hierarchy. The first stage of the hierarchy corresponds to the joint density function of the data y such that:

$$y \sim N\left(Z^*\theta, \Omega\right)$$

where $\Omega = E\left[uu'\right]$. The second stage of the hierarchy is defined as

$$\theta \sim N\left(\overline{\theta}, \Delta\right)$$

and the third stage of the hierarchy corresponds to the prior distribution of $\overline{\theta}$:

$$\overline{\theta} \sim N\left(\varphi, \Psi\right).$$

So, the marginal distribution of y conditional on $\overline{\theta}$ is

$$y \sim N\left(Z^*\overline{\theta}, \Sigma\right)$$

where $\Sigma = \Omega + Z^*\Delta Z^{*\prime}$. Assuming prior ignorance at the third stage of the hierarchy (i.e., $\Psi^{-1} = 0$) yields to the following posterior distribution of $\overline{\theta}$:

$$\overline{\theta} \sim N\left(\left(Z^{*\prime}\Sigma^{-1}Z^*\right)^{-1}Z^{*\prime}\Sigma^{-1}y, \left(Z^{*\prime}\Sigma^{-1}Z^*\right)^{-1}\right)$$

$$\overline{\theta} \sim N\left(\overline{\theta}^*, V^*\right).$$

Table 16.2 Shrinkage estimators

	Estimators of σ_i^2
Swamy	$\frac{1}{T-k}\left(y_i - Z_i\widehat{\theta}_{i,\mathrm{OLS}}\right)'\left(y_i - Z_i\widehat{\theta}_{i,\mathrm{OLS}}\right)$
ML	$\frac{1}{T}\left(y_i - Z_i\widehat{\theta}_i\right)'\left(y_i - Z_i\widehat{\theta}_i\right)$
Iterative Bayes	$\frac{1}{T+2}\left(y_i - Z_i\widehat{\theta}_i^*\right)'\left(y_i - Z_i\widehat{\theta}_i^*\right)$
Empirical Bayes	$\frac{1}{T-k}\left(y_i'y_i - y_i'Z_i\widehat{\theta}_{i,\mathrm{OLS}}\right)$
Empirical Iterative Bayes	$\frac{1}{T-k}\left(y_i - Z_i\widehat{\theta}_i^*\right)'\left(y_i - Z_i\widehat{\theta}_i^*\right)$
	Estimators of Δ
Swamy	$\widehat{\Delta} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\widehat{\theta}_{i,\mathrm{OLS}} - \frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}_{i,\mathrm{OLS}}\right)\left(\widehat{\theta}_{i,\mathrm{OLS}} - \frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}_{i,\mathrm{OLS}}\right)' - \frac{1}{N}\sum_{i=1}^{N}\widehat{\sigma}_i^2\left(Z_i'Z_i\right)^{-1}$
ML	$\widehat{\Delta} = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{\theta}_i - \widehat{\overline{\theta}}\right)'\left(\widehat{\theta}_i - \widehat{\overline{\theta}}\right)$
Iterative Bayes	$\widehat{\Delta}^* = \frac{1}{N-k-1}\left[R + \sum_{i=1}^{N}\left(\widehat{\theta}_i^* - \widehat{\overline{\theta}}^*\right)\left(\widehat{\theta}_i^* - \widehat{\overline{\theta}}^*\right)'\right]$
Empirical Bayes	$\widehat{\Delta}^* = \frac{1}{N-1}\left[\sum_{i=1}^{N}\left(\widehat{\theta}_{i,\mathrm{OLS}} - \widehat{\overline{\theta}}^*\right)\left(\widehat{\theta}_{i,\mathrm{OLS}} - \widehat{\overline{\theta}}^*\right)'\right]$
Empirical Iterative Bayes	$\widehat{\Delta}^* = \frac{1}{N-1}\left[R + \sum_{i=1}^{N}\left(\widehat{\theta}_i^* - \widehat{\overline{\theta}}^*\right)\left(\widehat{\theta}_i^* - \widehat{\overline{\theta}}^*\right)'\right]$
	Estimators of θ_i
Swamy	No estimator
ML	$\widehat{\theta}_i = \left[\widehat{\Delta}^{-1} + \frac{1}{\widehat{\sigma}_i^2}Z_i'Z_i\right]^{-1}\left[\widehat{\Delta}^{-1}\widehat{\overline{\theta}} + \frac{1}{\widehat{\sigma}_i^2}Z_i'Z_i\widehat{\theta}_{i,\mathrm{OLS}}\right]$
Bayes	$\widehat{\theta}_i^* = \left[\widehat{\Delta}^{*-1} + \frac{1}{\widehat{\sigma}_i^2}Z_i'Z_i\right]^{-1}\left[\widehat{\Delta}^{*-1}\widehat{\overline{\theta}}^* + \frac{1}{\widehat{\sigma}_i^2}Z_i'Z_i\widehat{\theta}_{i,\mathrm{OLS}}\right]$
	Estimators of $\overline{\theta}$
Swamy	$\widehat{\overline{\theta}} = \sum_{i=1}^{N}\Theta_i\widehat{\theta}_{i,\mathrm{OLS}}$
Empirical Bayes	$\widehat{\overline{\theta}}^* = \frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}_{i,\mathrm{OLS}}$
Others	$\widehat{\overline{\theta}}^*$ (and $\widehat{\overline{\theta}}$) $= \frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}_i^*$

Following (16.20), (16.21), and (16.22), we get:

$$\overline{\theta}^* = \sum_{i=1}^{N}\Theta_i\widehat{\theta}_{i,\mathrm{OLS}} \tag{16.37}$$

where

$$\Theta_i = \left(\sum_{i=1}^{N}\left[\Delta + \sigma_i^2\left(Z_i'Z_i\right)^{-1}\right]^{-1}\right)^{-1}\left[\Delta + \sigma_i^2\left(Z_i'Z_i\right)^{-1}\right]^{-1}.$$

The posterior mean $\overline{\theta}^*$ is a weighted average of the least squares estimates of individuals units and can be used as a point estimate of $\overline{\theta}$. In practice, the variance components Δ and σ_i^2 in (16.37) are unknown. The proposed solution of Lindley and Smith was an approximation that consists of using the mode of the posterior distribution rather than the mean. The proposed solution was labelled the empirical Bayes estimator. Normally, the marginal posterior densities of the parameters of interest can be obtained by integrating out the hyperparameters from the joint posterior density:

$$p(\theta|y) = \int p\left(\theta|\overline{\theta}, y\right) p\left(\overline{\theta}|y\right) \mathrm{d}\overline{\theta}.$$

The required integration poses an insurmountable challenge and closed-form analytic solutions cannot be obtained. Nevertheless, a full Bayesian implementation of the model is now feasible as a result of recent advances in sampling-based approaches to calculate marginal densities. The Gibbs sampling algorithm has been used successfully by Hsiao, Pesaran and Tahmiscioglu (1999).[6]

The Gibbs sampler is an iterative Markov Chain Monte Carlo (MCMC) method which only requires the knowledge of the full conditional densities of the parameter vector; see Chib (2001, 2007).

Starting from some arbitrary initial values, say $\theta^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_k^{(0)}\right)$ for a parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, it samples (generates) alternately from the conditional density of each component of the parameter vector conditional on the values of other components sampled in the last iteration M. The vectors $\left(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}\right)$ will form a Markov Chain with transition probability from stage θ' to the next one θ being:

$$H\left(\theta', \theta\right) = p\left(\theta_1|\theta_2', \theta_3', \ldots, \theta_k', y\right) p\left(\theta_2|\theta_1, \theta_3', \ldots, \theta_k', y\right) \ldots p\left(\theta_k|\theta_1, \ldots, \theta_{k-1}, y\right).$$

As the number of iterations $M \to \infty$, the samples values can be regarded as drawing from the true joint and marginal posterior densities. In order to implement Gibbs sampling, we need to specify the prior distribution of the hyperparameters. They are assumed to be independent and distributed as:

$$p\left(\Delta^{-1}, \Sigma\right) = W_k\left(\Delta^{-1}|\left(\rho R\right)^{-1}, \rho\right) \prod_{i=1}^{N} \sigma_i^{-2}$$

where W_k represents the Wishart distribution[7] with scale matrix (ρR) and degrees of freedom ρ. With this structure, the joint density of all the parameters may be written as:

---

[6] See also Hsiao (2003).

[7] A random symmetric positive definite (k,k) matrix A is said to follow a Wishart distribution $W_k(A|u,v)$ if the density of A is given by:

$$\frac{|A|^{(v-k-1)/2}}{|u|^{v/2}} \exp\left\{-\frac{1}{2} tr\left(u^{-1}A\right)\right\}.$$

See Koop (2003).

$$p\left(\theta_i,\overline{\theta},\Delta,\sigma_i^2|y,y_{i0}\right)\propto\prod_{i=1}^{N}\sigma_i^{-T}\exp\left[-\frac{1}{2}\sum_{i=1}^{N}\sigma_i^{-2}\left(y_i-Z_i\theta_i\right)'\left(y_i-Z_i\theta_i\right)\right]$$
$$\times|\Delta|^{-\frac{N}{2}}\exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\theta_i-\overline{\theta}\right)'\Delta^{-1}\left(\theta_i-\overline{\theta}\right)\right]$$
$$\times|\Psi|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left(\overline{\theta}-\varphi\right)'\Psi^{-1}\left(\overline{\theta}-\varphi\right)\right]$$
$$\times|\Delta|^{-\frac{1}{2}(\rho-k-1)}\exp\left[-\frac{1}{2}tr\left(\rho R\right)\Delta^{-1}\right]$$
$$\times\prod_{i=1}^{N}\sigma_i^{-2}.$$

The first line of the above formula corresponds to the standard likelihood function and the others represent the prior information. The relevant conditional distributions that are needed to implement the Gibbs sampler in this case are obtained from the joint posterior density:

$$p\left(\theta_i|y,\overline{\theta},\Delta^{-1},\sigma_1^2,\ldots,\sigma_N^2\right)=N\left[A_i\left(\sigma_i^{-2}Z_i'y_i+\Delta^{-1}\overline{\theta}\right),A_i\right]$$
$$p\left(\overline{\theta}|y,\theta_1,\ldots,\theta_N,\Delta^{-1},\sigma_1^2,\ldots,\sigma_N^2\right)=N\left[B\left(N\Delta^{-1}\widetilde{\theta}+\Psi^{-1}\varphi\right),B\right]$$
$$p\left(\Delta^{-1}|y,\theta_1,\ldots,\theta_N,\overline{\theta},\sigma_1^2,\ldots,\sigma_N^2\right)=W_k\left[\left(\sum_{i=1}^{N}\left(\theta_i-\overline{\theta}\right)\left(\theta_i-\overline{\theta}\right)'+\rho R\right)^{-1},\rho+N\right]$$
$$p\left(\sigma_i^2|y,\theta_1,\ldots,\theta_N,\overline{\theta},\Delta^{-1}\right)=\mathrm{IG}\left[T/2,\left(\left(y_i-Z_i\theta_i\right)'\left(y_i-Z_i\theta_i\right)\right)/2\right]$$

where

$$A_i=\left(\sigma_i^{-2}Z_i'Z_i+\Delta^{-1}\right)^{-1},B=\left(N\Delta^{-1}+\Psi^{-1}\right)^{-1},\widetilde{\theta}=\frac{1}{N}\sum_{i=1}^{N}\theta_i$$

for $i=1,\ldots,N$ and IG denotes the inverse gamma distribution. These values sampled after some initial number of iterations can be used to construct estimates of the parameters of interest. Hsiao, Pesaran and Tahmiscioglu (1999) call the estimator of $\overline{\theta}$ obtained using Gibbs sampling the "hierarchical Bayes" estimator.

16.3.3 An Example

Following Hsiao, Pesaran and Tahmiscioglu (1999), Baltagi, Bresson and Pirotte (2004) considered the simple dynamic version of the classical Tobin q investment model (16.13). The panel data set used in this study contains 337 U.S. firms over

17 years (1982–1998). The estimation results are given in Table 16.3. We first give the results of 9 homogeneous panel data estimators. These include OLS, which ignores the individual effects; the Within estimator, which allows for fixed individual effects; and FGLS, which assumes that individual effects are random. Hsiao, Pesaran and Tahmiscioglu (1999) report the fixed effects estimates for a subset of

Table 16.3 Estimates of the q investment model

Model type	Intercept	$(I/K)_{i,\,t-1}$	$q_{i,t}$
OLS	0.086	0.464	0.007
	(24.528)	(38.718)	(17.445)
Within	----	0.287	0.015
		(21.565)	(21.921)
FGLS	0.092	0.339	0.011
	(18.823)	(26.612)	(20.363)
2SLS	0.087	0.457	0.007
	(17.646)	(21.017)	(16.737)
Within-2SLS	----	0.163	0.016
		(5.429)	(22.176)
2SLS–KR	0.083	0.453	0.008
	(15.990)	(21.072)	(16.020)
FD2SLS	----	0.296	0.026
		(10.047)	(21.407)
FD2SLS–KR	----	0.313	0.022
		(16.419)	(20.648)
FDGMM	----	0.347	0.022
		(37.978)	(29.833)
Heterogeneous estimators			
Average OLS	0.019	0.277	0.042
	(2.557)	(20.254)	(17.011)
Swamy	0.041	0.327	0.028
	(4.915)	(20.357)	(9.655)
Average ML	0.050	0.313	0.026
	(20.483)	(55.947)	(34.101)
Average 2SLS	0.017	0.280	0.042
	(0.981)	(5.694)	(11.972)
Average empirical Bayes	0.037	0.297	0.032
	(8.148)	(29.185)	(24.557)
Average iterative empirical Bayes	0.054	0.332	0.023
	(25.059)	(57.470)	(34.983)
Average iterative Bayes	0.055	0.342	0.022
	(32.123)	(71.069)	(39.124)
Hierarchical Bayes	0.063	0.423	0.014
	(15.080)	(26.496)	(12.490)
Pooled mean group	0.039	0.374	0.022
	(10.197)	(20.613)	(34.416)

Note: Numbers in parentheses denote t-statistics.

273 firms over a longer time period 1973–1992. The estimate of the q coefficient is 0.009 compared to 0.015 for our sample. The corresponding estimates of the coefficient of lagged (I/K) are 0.295 and 0.287, respectively. The pooled OLS estimates yield a lower q coefficient estimate of 0.007 and a higher coefficient of lagged (I/K) of 0.464. The FGLS estimates lie in between the OLS and Within estimates.

Since our model is dynamic, we also focus on pooled estimators employing two-stage least squares (2SLS) using as instruments the exogenous variables and their lagged values. These 2SLS estimates were pretty close to OLS, while the Within 2SLS estimates yielded a lower estimate of the coefficient of lagged (I/K) than that of Within. In addition, we report the first-difference 2SLS (FD2SLS) estimator proposed by Anderson and Hsiao (1982) in which fixed or random individual effects are eliminated and predetermined variables are used as instruments. This yielded an even higher estimate of the q coefficient (0.026) than 2SLS but a lower estimate of the coefficient of lagged (I/K) of 0.296. Keane and Runkle (1992) (hereafter denoted by KR) suggest a modification of the 2SLS estimator that allows for any arbitrary type of serial correlation in the remainder error term. We refer to this estimator as 2SLS–KR. Still another variant of this estimator allows for any arbitrary form of serial correlation in the first differenced model. This is denoted as the FD2SLS–KR estimator. The 2SLS–KR estimates are close to those of 2SLS, while the FD2SLS–KR estimates are close to those of FD2SLS. Finally, following Arellano and Bond (1991), we used a generalized method of moments (GMM) estimator on the first-difference specification (FDGMM) with instruments in levels. This incorporates more orthogonality conditions than are usually used by the Anderson and Hsiao (1982) estimator as well as a general robust variance–covariance matrix specification allowed by GMM. This yielded estimates close to those of FD2SLS.

For the heterogeneous estimators of Table 16.3, we first compute individual OLS and 2SLS regressions. The average OLS and 2SLS estimates of the q coefficient are around 0.042, while the estimates of the coefficient of lagged (I/K) are around 0.28. These are higher for Tobin's q coefficient estimate and lower for the estimate of the coefficient of lagged (I/K) than the mean group estimator obtained by Hsiao, Pesaran and Tahmiscioglu (1999). The latter were 0.037 and 0.323, respectively. We also computed the Swamy (1970) random coefficient regression estimator which is a weighted average of the individual least squares estimates where the weights are inversely proportional to their variance–covariance matrices. This yielded a lower q coefficient estimate of 0.026 than average OLS and a higher estimate of the coefficient of lagged (I/K) of 0.327. From the individual Maximum Likelihood estimators, based on the normality assumption, several shrinkage estimators have been proposed in the literature including the empirical Bayes estimator, the iterative Bayes estimator, and the iterative empirical Bayes estimator. The average ML estimates are close to those of Swamy. The average empirical Bayes estimate of the q coefficient is 0.032 while that of average iterative empirical Bayes and average iterative Bayes are 0.023 and 0.022, respectively. Next, we compute the Hsiao, Pesaran and Tahmiscioglu (1999) hierarchical Bayes estimates. This yields a q coefficient estimate of 0.014 compared to 0.0174 for the different sample used by

Hsiao, Pesaran and Tahmiscioglu (1999). The corresponding estimates of the coefficient of lagged (I/K) are 0.423 and 0.431, respectively. Normal densities for lagged (I/K) and q coefficients are drawn. Finally, we compute the Pesaran, Shin and Smith (1999) Pooled Mean Group estimator. This estimator constrains the long-run coefficients to be identical but allows the short-run coefficients and error variances to differ across individuals. Long-run coefficients and individual-specific error correction coefficients are estimated using Maximum Likelihood. These ML estimates are referred to as pooled mean group estimators in order to highlight the pooling effect of the homogeneity restrictions on the estimates of the long-run coefficients and the fact that averages across individuals are used to obtain individual-wide mean estimates of the error-correction coefficients and the other short-run parameters of the model. This yields a q coefficient estimate of 0.022 and a lagged (I/K) coefficient estimate of 0.374.

16.4 Comments on the Predictive Approach

In some problems, it is interesting to predict one or more post-sample observations on a given individual over several periods. In the Bayesian context, the problem of prediction is solved by calculation of the predictive density.

16.4.1 From the Post-sample Predictive Density...

A fundamental goal in any statistical analysis is to predict a set of future observations at time $(T+\tau)$, say $Y_{T+\tau}$, given the observed data Y_T and the underlying model M. Forecasting in the Bayesian context is done through the calculation of the prediction density defined as the distribution $Y_{T+\tau}$ conditioned on (Y_T, M) but marginalized over the parameters θ. The post-sample predictive density is defined as (see Hsiao and Tahmiscioglu (1997)):

$$p(Y_{T+\tau}|Y_T, M) = \int p(Y_{T+\tau}|Y_T, M, \theta)\, p(\theta|Y_T, M)\, \mathrm{d}\theta$$

where $p(Y_{T+\tau}|Y_T, M, \theta)$ is the conditional density of $Y_{T+\tau}$ given (Y_T, M, θ) and the marginalization is with respect to the posterior density $p(\theta|Y_T, M)$ of θ. Thus, when interested in forecasting future values $Y_{T+\tau}$, one uses the posterior distribution $p(\theta|Y_T, M)$ to integrate out the parameters and gets the predictive density where $p(Y_{T+\tau}|Y_T, \theta, M)$ is obtained from the sampling model. The Bayesian approach naturally gives rise to predictive densities where all parameters are integrated out, making it a perfect tool for forecasting. Of course, all this comes at a cost, which is typically of a computational nature. In general, the predictive density is not available in closed form. An analytical solution to the computational problem is provided by summarizing prior information through restrictions to natural-conjugate prior

densities (see Zellner (1971)). A natural-conjugate prior shares the functional form of the likelihood. When it belongs to exponential families, this leads to posterior densities of the same form. But now, a full Bayesian implementation of the model is now feasible using sampling-based approaches to calculate marginal densities. Using Gibbs sampling, Hsiao, Pesaran and Tahmiscioglu (1999) have proposed the "hierarchical Bayes" estimator. For individual i, if we want to predict y at time $T+1$, say $y_{i,T+1}$, we should use the conditional density $p(y_{i,T+1}|Y_T)$ where Y_T includes all observed data $(y_{i,1},\ldots,y_{i,T})$ as well as the data on explanatory variables $\left(Z'_{i,1},\ldots,Z'_{i,T},Z'_{i,T+1}\right)'$ where $Z'_{i,T+1}$is the k-vector of future explanatory variables $\left(1,y_{i,T},X'_{i,T+1}\right)'$.[8]

Consequently, for the first step-ahead forecast, the predictive density is defined as:

$$p(y_{i,T+1}|Y_{i,T}) = \int p(y_{i,T+1}|Y_{i,T},\theta)\,p(\theta|Y_{i,T})\,\mathrm{d}\theta$$

$$\sim N\left(E(y_{i,T+1}|Y_{i,T});V(y_{i,T+1}|Y_{i,T})\right),\ Y'_{i,T} \equiv (y'_i \vdots y_{i,T+1})'.$$

So, the expected future value $\widehat{y}_{i,T+1}$ is the mean of draws from the normal distribution. Using the properties of the multivariate normal distribution, we define the conditional density of $y_{i,T+2}$ given the observed data and the parameters, and so on. So, differences—between forecast values at time $T+\tau$—for several Bayes estimators (Empirical Bayes, Iterative Bayes, Empirical Iterative Bayes and Hierarchical Bayes) come from the differences in the estimation of $\left(\overline{\theta},\Delta,\sigma_i^2\right)$.

16.4.2 ...to the Good Forecast Performance of the Hierarchical Bayes Estimator: An Example

Let us consider our simple dynamic version of the classical Tobin q investment model (16.13). For prediction comparison, Baltagi, Bresson and Pirotte (2004) have estimated the model using the observations from 1982 to 1993 and have reserved the last 5 years for obtaining forecasts (1994–1998). Table 16.4 gives a comparison of various predictors using the RMSE criterion for the q investment model. Because of the ability of an estimator to characterize long-run as well as short-run responses is at issue, the average RMSE is calculated across the 337 firms at different forecast horizons. Specifically, each model was applied to each firm, and out-of-sample forecasts for 5 years were calculated. The relative forecast rankings are reported in Table 16.4 after 1 and 5 years. The overall average ranking for the full 5 year period is also reported. A comparison of heterogeneous versus homogeneous estimators reveals some interesting patterns. The average OLS, average 2SLS, and the pooled mean group estimators perform poorly, ranking always in the bottom of Table 16.4 no matter what forecast horizon we look at. The Swamy random coefficients

[8] For more details, see Chib (2005).

Table 16.4 Comparison of forecast performance of the q investment model

	1st year		5th year		Five-year average	
Ranking	Estimator	RMSE[1]	Estimator	RMSE[1]	Estimator	RMSE[1]
1.	Hierarchical Bayes	6.6781	OLS	10.0769	Hierarchical Bayes	8.5307
2.	Individual ML	6.9151	2SLS–KR	10.0825	FGLS	8.8064
3.	Iterative Bayes	6.9651	2SLS	10.0915	Iterative empirical Bayes	8.8069
4.	Iterative empirical Bayes	7.0024	Hierarchical Bayes	10.1428	Iterative Bayes	8.8464
5.	FGLS	7.0722	FGLS	10.1968	OLS	8.8957
6.	Empirical Bayes	7.0805	Iterative empirical Bayes	10.4385	2SLS–KR	8.9089
7.	OLS	7.1541	Iterative Bayes	10.6349	2SLS	8.9239
8.	2SLS–KR	7.1773	Within	10.9203	Individual ML	8.9909
9.	2SLS	7.1970	Within-2SLS	10.9614	Empirical Bayes	9.2750
10.	FD2SLS	7.4861	Individual ML	10.9756	Within	9.2786
11.	FD2SLS–KR	7.5008	Empirical Bayes	11.4226	Within-2SLS	9.4586
12.	Within	7.5030	FD2SLS–KR	11.9677	FD2SLS–KR	9.9345
13.	FDGMM	7.6695	FD2SLS	12.0473	FD2SLS	9.9486
14.	Individual OLS	7.7484	FDGMM	12.5747	FDGMM	10.2930
15.	Within-2SLS	7.8644	Individual OLS	13.6907	Individual OLS	10.6765
16.	Individual 2SLS	8.5933	Swamy	16.1467	Swamy	14.0715
17.	Swamy	11.9773	Average OLS	19.833	Individual 2SLS	14.1792
18.	Pooled mean group	12.9823	Average 2SLS	21.8026	Average OLS	17.2825
19.	Average OLS	14.9043	Individual 2SLS	21.8941	Pooled mean group	17.4408
20.	Average 2SLS	15.5311	Pooled mean group	22.0320	Average 2SLS	18.6442

[1] RMSE $\times 10^{-2}$

estimator did not perform well either, having a rank of 17 or 16 depending on the forecast horizon. The weak forecast performance of the average and the Swamy estimators relative to the homogeneous estimators arises because of the parameter-instability problem of the individual firm regressions. The shrinkage iterative Bayes and iterative empirical Bayes estimators perform well, ranking 3, 4 in the first year, 7, 6 in the 5th year, and 4, 3 for the 5 year average. The overall RMSE forecast rankings offer a strong endorsement for the iterative shrinkage estimators. However, this good performance is closely matched by some of the homogeneous estimators: FGLS, OLS, 2SLS–KR and 2SLS. These rank $5,7,8,9$ in the first year, $5,1,2,3$ in the 5th year, and $2,5,6,7$ for the 5 year average. Hsiao, Pesaran and Tahmiscioglu (1999) also compared the out-of-sample forecasts for their sample using the fixed effects, the mean group, the corrected mean group, average empirical Bayes, and hierarchical Bayes for a 5 year horizon. The hierarchical Bayes estimator was found to perform the best for 1-to-5 year forecasts using the RMSE criterion. For our sample, our results confirm Hsiao, Pesaran and Tahmiscioglu (1999) conclusions. The hierarchical Bayes estimators ranks 1 for the first year, 4 for the 5th year, and 1 for the 5 year average. Its forecast performance is better than all other heterogeneous estimators and is better than usual homogeneous estimators (OLS, FGLS, 2SLS, ...) for the first year and for the 5 year average.

Baltagi, Bresson and Pirotte, (2004) reconsider the Tobin q investment model studied by Hsiao, Pesaran and Tahmiscioglu, (1999) using a panel of 337 U.S. firms over the period 1982–1998. It contrasts the out-of-sample forecast performance of 9 homogeneous panel data estimators and 11 heterogeneous and shrinkage Bayes estimators over a 5 year horizon. Results show that the average heterogeneous estimators perform the worst in terms of mean squared error, while the hierarchical Bayes estimator suggested by Hsiao, Pesaran and Tahmiscioglu (1999) performs the best. Homogeneous panel estimators and iterative Bayes estimators are a close second.

16.5 Conclusion

Although the performance of various estimators and their corresponding forecasts may vary in ranking from one empirical example to another (see Baltagi (1997)), Baltagi, Griffin and Xiong (2000), Baltagi, Bresson, Griffin and Pirotte (2003) and Baltagi, Bresson and Pirotte (2002, 2004), the consistent finding in all these studies is that homogeneous panel data estimators perform well in forecast performance mostly due to their simplicity, their parsimonious representation, and the stability of the parameter estimates. Average heterogeneous estimators perform badly due to parameter estimate instability caused by the estimation of several parameters with short time series. Shrinkage estimators did well for some applications, especially iterative Bayes and iterative empirical Bayes. For the Tobin q example, the hierarchical Bayes estimator performs very well and gives in mean the best forecasts.

References

Anderson, T.W. and C. Hsiao, 1982, Formulation and estimation of dynamic models using panel data, Journal of Econometrics 18, 47–82.

Arellano, M. and S. Bond, 1991, Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations, Review of Economic Studies 58, 277–297.

Baltagi, B., 1981, Pooling: An experimental study of alternative testing and estimation procedures in a two-way error components model, Journal of Econometrics 17, 21–49.

Baltagi, B., 2005, Econometric Analysis of Panel Data, Wiley, Chichester.

Baltagi, B.H. and J.M. Griffin, 1997, Pooled estimators vs. their heterogeneous counterparts in the context of dynamic demand for gasoline, Journal of Econometrics 77, 303–327.

Baltagi, B.H., J. Hidalgo, and Q. Li, 1996, A non-parametric test for poolability using panel data, Journal of Econometrics 75, 345–367.

Baltagi, B.H., J.M. Griffin, and W. Xiong, 2000, To pool or not to pool: Homogeneous versus heterogeneous estimators applied to cigarette demand, Review of Economics and Statistics 82, 117–126.

Baltagi, B.H., G. Bresson, and A. Pirotte, 2002, Comparison of forecast performance for homogeneous, heterogeneous and shrinkage estimators: Some empirical evidence from US electricity and natural-gas consumption, Economics Letters 76, 375–382.

Baltagi, B.H., G. Bresson, J.H. Griffin, and A. Pirotte, 2003, Homogeneous, heterogeneous or shrinkage estimators? Some empirical evidence from french regional gasoline consumption, Empirical Economics 28, 795–811.

Baltagi, B.H., G. Bresson, and A. Pirotte, 2004, Tobin q: Forecast performance for hierarchical Bayes, heterogeneous and homogeneous panel data estimators, Empirical Economics 29, 107–113.

Bun, M., 2004, Testing poolability in a system of dynamic regressions with nonspherical disturbances, Empirical Economics 29, 89–106.

Chib, S., 2001, Markov chain Monte Carlo methods: Computation and inference, in Handbook of Econometrics, J. Heckman and E. Leamer, eds., vol. 5, Chap. 57, North-Holland, Amsterdam, 3570–3649.

Chib, S., 2007, Panel data modeling and inference: A bayesian primer, in The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice, L. Màtyàs and P. Sevestre, eds., Chap. 14, Kluwer Academic Publishers, Dordrecht (forthcoming).

Driver, C., K. Imai, P. Temple, and A. Urga, 2004, The effect of uncertainty on UK investment authorisation: Homogeneous vs. heterogeneous estimators, Empirical Economics 29, 115–128.

Gelfand, A.E. and A.F.M. Smith, 1990, Sampling-based approaches to marginal densities, Journal of the American Statistical Association 46, 84–88.

Gelfand, A.E. and A.F.M. Smith, 1992, Bayesian statistics without tears: A sampling–resampling perspective, American Statistician 46, 84–88.

Hsiao, C., 2003, Analysis of Panel Data, Second Edition, Cambridge University Press, Cambridge.

Hsiao, C. and H.M. Pesaran, 2007, Random coefficient panel data models, in The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice, L. Màtyàs and P. Sevestre, eds., Chap. 5, Kluwer Academic Publishers, Dordrecht (forthcoming).

Hsiao, C. and A.K. Tahmiscioglu, 1997, A panel analysis of liquidity constraints and firm investment, Journal of the American Statistical Association 92, 455–465.

Hsiao, C., M.H. Pesaran, and A.K. Tahmiscioglu, 1999, Bayes estimation of short-run coefficients in dynamic panel data models, in Analysis of Panels and Limited Dependent Variable Models, C. Hsiao, K. Lahiri, L.-F. Lee, and M.H. Pesaran, eds., Cambridge University Press, Cambridge, 268–296.

Judge, G.G. and M.E. Bock, 1978, The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics, North-Holland, Amsterdam.

Keane, M.P. and D.E. Runkle, 1992, On the estimation of panel-data models with serial correlation when instruments are not strictly exogenous, Journal of Business and Economic Statistics 10, 1–9.

Koop, G., 2003, Bayesian Econometrics, Wiley, Chichester.

Lindley, D.V. and A.F.M. Smith, 1972, Bayes estimates for the linear model, Journal of the Royal Statistical Society B, 34, 1–41.

Maddala, G.S., R.P. Trost, H. Li, and F. Joutz, 1997, Estimation of short-run and long-run elasticities of energy demand from panel data using shrinkage estimators, Journal of Business and Economic Statistics 15, 90–100.

Maddala, G.S. and W. Hu, 1996, The pooling problem, in The Econometrics of Panel Data: A Handbook of Theory with Applications, L. Màtyàs and P. Sevestre, eds., Kluwer Academic Publishers, Dordrecht, 307–322.

McElroy, M.B., 1977, Weaker MSE criteria and tests for linear restrictions in regression models with non-spherical disturbances, Journal of Econometrics 6, 389–394.

Pesaran, M.H. and R. Smith, 1995, Estimating long-run relationships from dynamic heterogenous panels, Journal of Econometrics 68, 79–113.

Pesaran, M.H. and Z. Zhao, 1999, Bias reduction in estimating long-run relationships from dynamic heterogeneous panels, in Analysis of Panels and Limited Dependent Variable Models, C. Hsiao, K. Lahiri, L.-F. Lee and M.H. Pesaran, eds., Cambridge University Press, Cambridge, 297–322.

Pesaran, M.H., R. Smith, and K.S. Im, 1996, Dynamic linear models for heterogenous panels, in The Econometrics of Panel Data: A Handbook of Theory with Applications, L. Màtyàs and P. Sevestre, eds., Kluwer Academic Publishers, Dordrecht, 145–195.

Pesaran, M.H., Y. Shin, and R. Smith, 1999, Pooled mean group estimation of dynamic heterogeneous panels, Journal of the American Statistical Association 94, 621–634.

Rao, C.R., 1975, Simultaneous estimation of parameters in different linear models and applications to biometric problems, Biometrics 31, 545–554.

Robertson, D. and J. Symons, 1992, Some strange properties of panel data estimators, Journal of Applied Econometrics 7, 175–189.

Roy, S.N., 1957, Some Aspects of Multivariate Analysis, Wiley, New York.

Smith, A.F.M., 1973, A general bayesian linear model, Journal of the Royal Statistical Society B, 35, 67–75.

Swamy, P.A.V.B., 1970, Efficient inference in a random coefficient regression model, Econometrica 38, 311–323.

Wallace, T.D., 1972, Weaker criteria and tests for linear restrictions in regression, Econometrica 40, 689–698.

Zellner, A., 1962, An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, Journal of the American Statistical Association 57, 348–368.

Zellner, A., 1971, An Introduction to Bayesian Inference in Econometrics, Wiley, New York.

Ziemer, R.F. and M.E. Wetzstein, 1983, A Stein-rule method for pooling data, Economics Letters 11, 137–143.

Chapter 17
Duration Models and Point Processes

Jean-Pierre Florens, Denis Fougère and Michel Mouchart

Many economic phenomena are characterized by the observation of a sequence of events on a continuous interval of time. Think, for instance, to observing the dates of a specific type of financial transactions, or to observing the dates of changes of the individual labor market situation (full-time employed, part-time employed, unemployed, etc.). The length of the interval between two successive events is called a duration. A duration is a positive random variable, denoted T, representing the length of a time period spent by an individual or a firm in a given state. For simplicity, we assume that the distribution of T is not defective, i.e. $\Pr(T = \infty) = 0$. This variable is also called a failure time when the date of change is interpreted as a breakdown or a failure.

The most elementary duration model is based on a "death process" $\{X_t, t \in \mathbb{R}_+\}$, for which X_t takes its values in the discrete state space $\{E_0, E_1\}$. At the time origin, called the birth date, the process is in state E_0, i.e. $X_0 = E_0$. Trajectories of the process X_t have at most a unique transition from state E_0 to state E_1, which occurs at time T, called the death date. Consequently, the duration T generated by a trajectory of the death process X_t is defined as follows:

$$T = \inf\{t \mid X_t = E_1\} .$$

In most structural models, T is a continuous random variable, *but* the empirical distribution function is a discrete time process and nonparametric methods are of-

Jean-Pierre Florens
Toulouse School of Economics, Institut Universitaire de France, Toulouse, France,
e-mail: florens@cict.fr

Denis Fougère
CNRS, CREST-INSEE (Paris), CEPR (London) and IZA (Bonn), e-mail: fougere@ensae.fr

Michel Mouchart
Emeritus Professor of Statistics and Econometrics, Institut de statistique, 20 Voie du Roman Pays, B-1348 Louvain-La-Neuve (Belgium), e-mail: Michel.Mouchart@uclouvain.be

This chapter summarizes and updates Chap. 9 (on Duration Models) and Chap. 20 (on Point Processes) in Mátyás and Sevestre (1996). Being a survey, this chapter by-passes many details, explanations and justifications, given in Mouchart (2004) in a textbook form, with a similar approach and notation.

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

ten based on (functional) transformations of the empirical distribution function, considered as the best estimator of the "true" distribution function. Therefore, in this chapter, we explicitly consider both continuous and discrete durations.

The first section of this survey concentrates on marginal models of durations, i.e. models without explanatory variables. It presents the main functions characterizing the distribution of a duration variable, the survivor and hazard functions among others. Section 17.2 is devoted to the presentation of conditional duration models, and more particularly, proportional hazards and accelerated life models, which incorporate the effects of explanatory variables in two different ways. In this section, a special emphasis is put on the problem of unobserved individual heterogeneity. The basic duration model treats a single spell (of unemployment, for example) ending with a given kind of transition (from unemployment to employment, for example).[1] But, in general, as a death could be due to various causes, an individual could exit from unemployment to enter one among different states: full-time employment, part-time employment, or training, for example. When a single-spell duration has many (at least two) outcomes, the duration model may be modelled by means of a so-called competing risks model. Competing risks models are presented in the third section, which also contains a discussion on their identifiability. The right-censoring issue is presented here as a particular case of a competing risks duration model. The fourth section is concerned with statistical inference, with a special emphasis on non- and semi- parametric estimation of single-spell duration models.

The remaining part of this chapter is devoted to point processes, which can be viewed as a generalization of duration models. Such processes are a mathematical formalization which allows to examine individual mobilities or transitions between a finite number of discrete states through (continuous) time. They are particularly useful for the microeconometric analysis of labor market dynamics. Section 17.5 sets forth the main definitions for point and counting processes. Distribution, intensity and likelihood functions of such processes are also examined. Section 17.6 presents important elementary examples of point processes, namely Poisson, Markov and semi-Markov processes. Such processes are of great interest because they are well adapted to the case of observed censored or truncated realizations. The last section presents a general semiparametric framework for studying point processes with explanatory variables. It also focuses on the definition of martingale estimators, which are particularly useful in this framework.

17.1 Marginal Duration Models

17.1.1 Distribution, Survivor and Density Functions

We first recall the general definition of the distribution function and of its complement, the survivor function. Next, we give more details for the continuous and

[1] Recently, duration models have been used to analyze the determinants of time intervals between two successive changes in the price of a product sold in a given outlet (see, for instance, Fougère, Le Bihan and Sevestre (2007)).

the discrete cases, particularly from the point of view of the continuity of these functions.

Definition 17.1. (Distribution function) The distribution function of the duration variable T is denoted F and is defined as

$$F(t) = \Pr(T \leq t), \quad t \geq 0 \,.$$

■

The main properties of the distribution function F are: $F(t) \in [0,1]$, F is monotone non-decreasing, right continuous and $\lim_{t\to\infty} F(t) = 1$.

Definition 17.2. (Survivor function) The survivor function of the duration variable T, denoted S, is defined as

$$S(t) = \Pr(T \geq t) = 1 - F(t) + \Pr(T = t) \,.$$

■

Its main properties are: $S(t) \in [0,1]$, S is monotone non-increasing, left-continuous and $\lim_{t\to\infty} S(t) = 0$.

Definition 17.3. (Density function) If there exists a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$F(t) = \int_0^t f(u)\mathrm{d}u \quad \text{or} \quad f(t) = \frac{dF}{\mathrm{d}t} = -\frac{dS}{\mathrm{d}t} \,,$$

f is called the density of T. ■

Thus, the density function may be interpreted as the "instantaneous probability" of a failure, a death or an exit (from unemployment, for instance). Remember that in the continuous case, there exists a value of t such that $F(t) = S(t) = 0.5$; that value is the median of the distribution.

Definition 17.4. (Discrete duration)

$$\exists (f_j, a_j), \quad j \in J \subseteq \mathbb{N},\ f_j > 0, \quad \sum\nolimits_{j\in J} f_j = 1,\ 0 \leq a_j < a_{j+1}$$

such that

$$F(t) = \sum\nolimits_{j\in J} f_j \, \mathbf{1}\left\{t \geq a_j\right\} = \sum\nolimits_{\left\{j | a_j \leq t\right\}} f_j$$

$$S(t) = \sum\nolimits_{j\in J} f_j \, \mathbf{1}\left\{t \leq a_j\right\} = \sum\nolimits_{\left\{j | a_j \geq t\right\}} f_j$$

or equivalently

$$\begin{aligned} f_j &= F\left(a_j\right) - F\left(a_{j^-}\right) = F\left(a_j\right) - F\left(a_{j-1}\right) \\ &= S\left(a_j\right) - S\left(a_{j^+}\right) \ = S\left(a_j\right) - S\left(a_{j+1}\right) \end{aligned}$$

■

In the framework of a death process, the event $\{T = a_j\}$ means "alive up to age a_j and dead at age a_j" and that event has probability f_j.

17.1.2 Truncated Distributions and Hazard Functions

The use of statistical duration models may be justified by several arguments:

(i) *Problem of time dependence.* Consider the following question. What is the "instantaneous" probability of dying at time t given you are still living at time t? More generally, this is the problem of the probability law of duration T, *conditional* on $T \geq t$ (remember that the event $\{T \geq t\}$ means "still alive at time t"). This problem is exactly that of analyzing the dynamic behavior of the process. Such conditional distributions are "truncated" distributions.

(ii) The preceding question is often so natural that modelling those truncated distributions may be *economically more meaningful* than modelling the untruncated distributions. For instance, in job search models, the reservation wage, at a given instant, is a function of the duration of unemployment up to that instant.

(iii) *Right-censoring* (see Sect. 17.3.4) makes truncated distributions particularly useful.

Definition 17.5. (Integrated hazard function) The integrated hazard function of the duration variable T is denoted Λ and is defined as

$$\begin{aligned} \Lambda : \mathbb{R}_+ &\to \mathbb{R}_+ \\ t &\mapsto \Lambda(t) = \int_{[0,t[} \frac{1}{S(u)} \mathrm{d}F(u) \end{aligned}$$

■

The function Λ is monotone non-decreasing, left-continuous and verifies $\Lambda(0) = 0$ and $\Lambda(\infty) = \infty$. As we will see later, the integrated hazard function is a useful tool for characterizing some duration distributions. Let us consider now the hazard function (or age-specific failure rate).

(i) *Continuous case*

In the continuous case, there is a density function $f(t)$ and

$$\Lambda(t) = \int_0^t \frac{f(u)}{S(u)} \mathrm{d}u = -\int_0^t \frac{1}{S(u)} \mathrm{d}S(u) = -\ln S(t) \,.$$

Definition 17.6. (Hazard function) The hazard function of the duration variable T is denoted λ and is defined as

$$\lambda(t) = \mathrm{d}\Lambda(t)/\mathrm{d}t = f(t)/S(t) = -\mathrm{d} \ln S(t)/\mathrm{d}t \,.$$

■

The function $\lambda(t)$ may be viewed as the "instantaneous probability" of leaving the current state, indeed

$$\lambda(t) = \lim_{\Delta \to 0} \frac{\Pr[t \leq T < t + \Delta \mid T \geq t]}{\Delta} .$$

Thus, $\lambda(t)$ is also called the "age-specific failure rate" or the "age-specific death rate". The function λ is non negative and $\int_0^t \lambda(u)\,\mathrm{d}u < \infty$, $\forall t \in R_+$, but $\int_0^\infty \lambda(u)\,\mathrm{d}u = \infty$ for non-defective distributions. Note that λ is *not* necessarily monotone.

Straightforward relationships between the distribution, survivor and hazard functions should be noticed:

$$\Lambda(t) = \int_0^t \lambda(u)\,\mathrm{d}u,\ f(t) = \lambda(t) \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right)$$

$$S(t) = \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right),\ F(t) = 1 - \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right)$$

which shows that each of these functions completely characterizes the distribution of a duration.

Definition 17.7. (Temporal independence) The hazard function of the duration T has the property of *temporal independence* if and only if it is constant over time, i.e. $\lambda(t) = \lambda,\ \forall t \in \mathbb{R}\ (\lambda > 0)$ ■

(ii) *Discrete case*

Remember that, in the discrete case, for any (integrable) function $g(u)$ we have

$$\int_{[0,t[} g(u)\mathrm{d}F(u) = \sum_{\{j|a_j<t\}} g(a_j) f_j = \sum_j g(a_j) f_j \mathbf{1}\{a_j < t\} .$$

Therefore

$$\Lambda(t) = \sum_{\{j|a_j<t\}} \frac{f_j}{S(a_j)} = \sum_{\{j|a_j<t\}} \frac{f_j}{f_j + f_{j+1} + \ldots} .$$

So, we obtain the discrete version of the (instantaneous) hazard function as

$$\lambda_j = \Lambda(a_{j^+}) - \Lambda(a_j) = \frac{f_j}{f_j + f_{j+1} + f_{j+2} + \ldots} = \frac{f_j}{S(a_j)} .$$

In particular, $\lambda_1 = f_1$. The last formula may also be interpreted as

$$\lambda_j = \Pr(T = a_j \mid T \geq a_j) .$$

To deduce relationships between survivor and hazard functions in the discrete case, let us write the survivor function as:

$$S(t) = \prod\nolimits_{\{j|a_j<t\}} (1 - \lambda_j)$$

based on the familiar identity

$$a_0 + a_1 = a_0\left(1 + \frac{a_1}{a_0}\right)$$

$$a_0 + a_1 + a_2 = a_0\left(1 + \frac{a_1}{a_0}\right)\left(1 + \frac{a_2}{a_0 + a_1}\right)$$

$$\cdots$$

$$\Sigma_{0\le j<k}\, a_j = a_0 \prod_{1\le j<k}\left(1 + \frac{a_j}{\Sigma_{0\le m<j-1}\, a_m}\right)$$

applied to:

$$S(t) = 1 - \sum_{\{j|a_j<t\}} \lambda_j \,.$$

Thus we obtain the relationship

$$\ln S(t) = \Sigma_{\{j|a_j<t\}} \ln(1-\lambda_j) \approx -\Sigma_{\{j|a_j<t\}}\, \lambda_j = -\Lambda(t)$$

if λ_j is "small", i.e. $-\ln\,(1-\lambda_j) \approx \lambda_j$. Thus, in the discrete case, $\Lambda(t)$ is *approximately* equal to $-\ln S(t)$ if all λ_j are small, while in the continuous case, $\Lambda(t)$ is exactly equal to $-\ln S(t)$. Moreover, in the discrete case:

$$f_j = \lambda_j \prod_{1\le i\le j-1}(1-\lambda_i) \,.$$

Figure 17.1 presents the main distributions used for the statistical analysis of duration data.

17.2 Conditional Models

17.2.1 General Considerations

17.2.1.1 The Two Levels of Analysis to be Considered

(i) For a *descriptive* (*or exploratory*) *data analysis*, covariates may be used to control for observable factors of heterogeneity by performing separate analyses.
(ii) When the objective is to estimate a *structural model*, the parameter of interest may be such that the (marginal) process generating some covariates may be uninformative about the parameter of interest which, at the same time, is a function of a parameter sufficient to parametrize the process conditional on those covariates. Those covariates are then called "exogenous variables" and are generally denoted by Z whereas the other variables, denoted by Y (or T, in case of a duration variable), are called "endogenous", because the model describes the way they are generated conditionally on the exogenous variables.

Distribution	Parameters	Survivor function $S(t)$	Density function $f(t)$	Hazard function $h(t)$	Time variation of the hazard function $\mathrm{d}h(t)/\mathrm{d}t$
Exponential	$\lambda > 0$	$\exp(-\lambda t)$	$\lambda \exp(-\lambda t)$	λ	constant $(\mathrm{d}h(t)/\mathrm{d}t=0)$
Weibull	$\alpha > 0, \lambda > 0$	$\exp(-\lambda t^{\alpha})$	$\alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha})$	$\alpha\lambda t^{\alpha-1}$	increasing if $\alpha>1$ decreasing if $\alpha<1$ constant if $\alpha=1$ $(h(t)=\lambda)$
Gamma	$\lambda>0,\ \gamma>0$	$\frac{1}{\Gamma(\gamma)}\int_t^{\infty}\lambda(\lambda s)^{\gamma-1}\exp(-\lambda s)\mathrm{d}s$	$\frac{\lambda(\lambda t)^{\gamma-1}\exp(-\lambda t)}{\Gamma(\gamma)}$	$\frac{\lambda(\lambda t)^{\gamma-1}\exp(-\lambda t)}{\int_t^{\infty}\lambda(\lambda s)^{\gamma-1}\exp(-\lambda s)\mathrm{d}s}$	increasing if $\gamma>1$, decreasing if $\gamma<1$, constant if $\gamma=1$ $(h(t)=\lambda)$
Generalized Gamma	$\lambda>0,\ \alpha>0,\ \gamma>0$	$\frac{1}{\Gamma(\gamma)}\int_t^{\infty}\lambda\alpha(\lambda s)^{\alpha\gamma-1}\exp(-(\lambda s)^{\alpha})\mathrm{d}s$	$\frac{\lambda\alpha(\lambda t)^{\alpha\gamma-1}\exp(-(\lambda t)^{\alpha})}{\Gamma(\gamma)}$	$\frac{\lambda\alpha(\lambda t)^{\alpha\gamma-1}\exp(-(\lambda t)^{\alpha})}{\int_t^{\infty}\lambda\alpha(\lambda s)^{\alpha\gamma-1}\exp(-(\lambda s)^{\alpha})\mathrm{d}s}$	Gamma distribution with parameter $\lambda>0$ *et* $\gamma>0$ *if* $\alpha=1$, exponential distribution with parameter λ if $\alpha=1$ and $\gamma=1$
Lognormal	$\mu, \sigma > 0$	$1-\Phi\left[\frac{\ln t-\mu}{\sigma}\right]$	$\frac{1}{\sigma t}\varphi\left[\frac{\ln t-\mu}{\sigma}\right]$	$\frac{\frac{1}{\sigma t}\varphi\left[\frac{\ln t-\mu}{\sigma}\right]}{1-\Phi\left[\frac{\ln t-\mu}{\sigma}\right]}$	increasing, then decreasing
Log-logistic	$\lambda > 0, \gamma > 0$	$\left[1+(\lambda t)^{\gamma}\right]^{-1}$	$\gamma\lambda^{\gamma}t^{\gamma-1}[1+(\lambda t)^{\gamma}]^{-2}$	$\gamma\lambda^{\gamma}t^{\gamma-1}[1+(\lambda t)^{\gamma}]^{-1}$	increasing then decreasing if $\gamma>1$, decreasing if $\gamma\leq 1$
Singh-Maddala	$\lambda > 0, \alpha > 0, \gamma > 0$	$\left[1+(\lambda t)^{\gamma}\right]^{-\left(\frac{\alpha}{\lambda}\right)^{\gamma}}$	$\gamma\alpha^{\gamma}t^{\gamma-1}\left[1+(\lambda t)^{\gamma}\right]^{-2}$	$\gamma\lambda^{\gamma}t^{\gamma-1}\left[1+(\lambda t)^{\gamma}\right]^{-1}$	log-logistic (λ, γ) if $\lambda=\alpha$, Weibull $(\alpha^{\gamma}, \gamma)$ if $\lambda=0$, exponential (α) if $\lambda=0$ and $\gamma=1$, increasing then decreasing if $\gamma>1$, decreasing if $\gamma<1$

Fig. 17.1 Examples of distributions for durations

In such a case, it is admissible to specify only the process conditional on those exogenous variables, leaving the marginal process generating those exogenous variables virtually unspecified. In other words, for the parameter of interest, $p(t \mid z, \theta)$ is as informative as $p(t, z \mid \theta)$. According to a general principle of parsimony, the conditional model is therefore preferred.

17.2.1.2 How to Specify conditional Models

(i) In general, a natural way of specifying conditional models is to make the parameters of a distribution dependent on the conditioning variable. Thus, in $F_T(t \mid \theta)$, one would transform θ into $g(z, \theta)$ where g would be a known function. For example, $Y \sim N(\mu, \sigma^2)$ could be transformed into $(Y \mid Z) \sim N(\alpha + \beta Z, \sigma^2)$. Similarly, $T \sim \exp(\theta)$ could be transformed into $(T \mid Z) \sim \exp[g(Z, \theta)]$ where, e.g. $g(Z, \theta) = \exp(-Z'\theta)$.

(ii) When modelling individual data (and, in particular, duration data), a frequently used strategy consists of starting with a so-called "baseline" distribution for a reference individual, i.e. either an individual not belonging to the treatment group (e.g. an individual for which $Z = 0$) or a "representative" individual (e.g. an individual for which $Z = E(Z)$) and thereafter modelling, what makes the other individuals different from that individual of reference. Typical examples are the following:

- in the proportional hazard model, the global effect of all regressors Z is to multiply the baseline hazard function by a scale factor,
- in the accelerated life model, the global effect of all regressors Z is to rescale the duration variable. From now on, we shall only use the notation θ for the complete parameter characterizing the conditional distribution generating $(T \mid Z)$. This vector is decomposed into $\theta = (\alpha, \beta)$ where α parametrizes the baseline distribution and β represents the effect of the exogenous variables.

17.2.1.3 Time-Varying and Time-Constant Covariates Must be Distinguished

The covariates may represent:

- individual characteristics, such as gender, level of education, and so on, which are fixed over time,
- other individual characteristics, such as marital status, number of children, eligibility to social benefits or programs, which are typically varying through time,
- but also characteristics of the macroeconomic environment, such as the unemployment rate, the job vacancy rate, the employment structure, and so on, which are also time-varying but possibly common to several individuals.

Some variables may also represent interactions between several covariates. The dynamic properties of the model and the estimation procedures crucially depends on whether the covariates are time-dependent or not.

17.2.1.4 Interpretation of the Parameters

Most models are typically nonlinear in the sense that partial derivatives (of interest) are not constant, but are functions of the values of the covariates and/or of the duration. This feature clearly makes the interpretation of the coefficients more difficult. Furthermore, those partial derivatives are often not those of conditional expectations (as in regression analysis) but those of hazard functions (i.e. of "instantaneous probabilities").

17.2.2 The Proportional Hazard or Cox Model

17.2.2.1 Definition

In the proportional hazard model, the effect of the exogenous variable is specified as multiplying a baseline hazard function by a function that depends on the exogenous variable. When Z is not time-dependent, this model is defined as

$$\lambda_T(t \mid z, \theta) = \lambda_0(t \mid \alpha) g(z, \beta), \quad \theta = (\alpha, \beta)\,,$$

where $\lambda_0(t \mid \alpha)$ is the so-called baseline hazard function and g is a known function. The proportional hazard model is equivalently characterized as

$$\Lambda_T(t \mid z, \theta) = g(z, \beta) \int_0^t \lambda_0(u \mid \alpha) \mathrm{d}u = g(z, \beta)\ \Lambda_0(t \mid \alpha)\,,$$

$$\begin{aligned} S_T(t \mid z, \theta) &= \exp\left\{-g(z, \beta) \textstyle\int_0^t \lambda_0(u \mid \alpha) \mathrm{d}u\right\} \\ &= \exp\left\{-g(z, \beta)\ \Lambda_0(t \mid \alpha)\right\} \\ &= [S_0(t \mid \alpha)]^{g(z, \beta)} \end{aligned}$$

where Λ_0 and S_0 are implicitly defined. Thus

$$\begin{aligned} f_T(t \mid z, \theta) &= \lambda_T(t \mid z, \theta)\ S_T(t \mid z, \theta) \\ &= g(z, \beta)\ \lambda_0(t \mid \alpha)\ [S_0(t \mid \alpha)]^{g(z, \beta)} \end{aligned}$$

17.2.2.2 Identification

The problem of identifying separately the functions g and λ_0 comes from the fact that for any $k > 0 : g \cdot \lambda_0 = gk \cdot k^{-1} \lambda_0$. A rather natural solution consists of defining a reference individual, i.e. a particular value z_0 of Z for which $g(z_0, \beta) = 1, \forall \beta$. Consequently, $\lambda_T(t \mid z_0, \theta) = \lambda_0(t \mid \alpha)$. When $Z = 0$ is meaningful, a typical normalization is $g(0, \beta) = 1$.

In the proportional hazard model with time-constant covariates, the first-order derivative

$$\frac{\partial}{\partial z}\ln \lambda_T(t \mid z,\theta) = \frac{\partial}{\partial z}\ln g(z,\beta)\ ,$$

depends on z and β only and is therefore independent of t.

17.2.2.3 Semi-parametric Modelling

When interest is focused on the role of the exogenous variables, α is treated as a nuisance parameter and β is the sole parameter of interest. In such a case, modelling often relies on one of the following two extreme possibilities:

(i) $\lambda_0(t \mid \alpha)$ is specified in the most simplest way such as $\lambda_0(t \mid \alpha) = \lambda_0(t)$, i.e. is completely known, or $\lambda_0(t \mid \alpha) = \alpha$, i.e. the baseline distribution is exponential and therefore depends on only one unknown parameter;

(ii) $\lambda_0(t \mid \alpha)$ is specified in the most general way: $\lambda_0(t \mid \alpha) = \alpha(t)$, i.e. a functional parameter (α is a non-negative function such that its integral on the positive real line diverges). This is a semiparametric model with parameter $\theta = (\alpha, \beta)$, where α takes its value in a functional space, whereas β takes its value in a (finite dimensional) Euclidean space. This approach is particularly attractive in situations where economic theory would not give much information on the structure of $\lambda_0(t \mid \alpha)$.

17.2.2.4 A Particular Case

The function $g(z,\beta)$ should clearly be non-negative. An easy way to obtain that property without restriction on β is the log-linear specification, viz.:

$$g(z,\beta) = \exp(z'\beta), \quad \beta \in \mathbb{R}^k\ .$$

In such a case $\Lambda_0(t|\alpha) = \Lambda_T(t|0,\theta)$. That specification has a number of interesting properties. First, let us remark that:

$$\frac{\partial}{\partial z}\ln \lambda_T(t \mid z,\theta) = \frac{\partial}{\partial z}\ln g(z,\beta) = \beta\ ,$$

i.e. z has a constant proportional effect on the instantaneous conditional probability of leaving state E_0. If z is not time-dependent, one may also write

$$S_T(t \mid z,\theta) = \exp\{-\Lambda_0(t \mid \alpha)\exp(z'\beta)\} = [S_0(t \mid \alpha)]^{\exp(z'\beta)}$$

$$f_T(t \mid z,\theta) = \lambda_0(t \mid \alpha)\exp(z'\beta)\ [S_0(t \mid \alpha)]^{\exp(z'\beta)}$$

Let us define

$$\varepsilon_t = -\ln \Lambda_0(t \mid \alpha) - z'\beta\ .$$

where ε_t has a completely specified distribution, independent of α, z or β, namely a unit double-exponential distribution. Then we may write

$$-\ln \Lambda_0(t \mid \alpha) = z'\beta + \varepsilon_t \ .$$

This is a (non-normal) *nonlinear* regression but linear if α is known. This feature of the proportional hazard model was used by Han and Hausman (1990) for conducting a semiparametric estimation on grouped duration data.

17.2.3 The Accelerated Time Model

17.2.3.1 The Basic Idea

In the accelerated time model, the effect of the exogenous variable is specified as modifying the time scale. For the ease of exposition, we assume that the exogenous variables are not time-dependent. The accelerated time model is accordingly defined as

$$T = [g(z,\beta)]^{-1} T_0 \quad \text{or} \quad T_0 = g(z,\beta)T$$

or, equivalently,

$$\begin{aligned}
\lambda_T(t \mid z,\theta) &= g(z,\beta) \times \lambda_0[t\ g(z,\beta) \mid \alpha] \\
\Lambda_T(t \mid z,\theta) &= \Lambda_0[t\ g(z,\beta) \mid \alpha] \\
S_T(t \mid z,\theta) &= S_0[t\ g(z,\beta) \mid \alpha] \\
f_T(t \mid z,\theta) &= g(z,\beta)\ f_0[t\ g(z,\beta) \mid \alpha]
\end{aligned}$$

with, as usual, $\theta = (\alpha,\beta)$. This specification may be particularly attractive when the baseline distribution admits a scale parameter.

17.2.3.2 Empirical Test for the Accelerated Time Model

Let us consider the quantile functions, i.e. the inverse of the survivor (rather than, as more usually, the distribution) functions:

$$q_T(p \mid z,\theta) = S_T^{-1}(p \mid z,\theta)\ ,\ 0 \le p \le 1\ ,$$

$$q_0(p \mid \alpha) = S_0^{-1}(p \mid \alpha), \quad 0 \le p \le 1\ .$$

Because of the strict monotonicity (in the continuous case) of the survivor function, we have

$$q_0(p \mid \alpha) = g(z,\beta) \cdot q_T(p \mid z,\theta)\ .$$

In the $\{q_0(p \mid \alpha), q_T(p \mid z,\theta)\}$–space, this gives, for a fixed value of z, an homogenous straight line, the gradient of which is given by $g(z,\beta)$. This feature suggests that an easy empirical test for the accelerated time model may be obtained through

an examination of the so-called "Q–Q-plot" (i.e. plot of the two quantiles) for a fixed value of Z and a fixed (typically, estimated) value of $\theta = (\alpha, \beta)$.

17.2.3.3 Regression Representation of the Accelerated Time Model

The accelerated time model may also be written, in logarithmic terms, as

$$\ln T = \ln T_0 - \ln g(z, \beta) .$$

If we define $\mu_0 = E[\ln T_0]$ and $\varepsilon = \ln T_0 - E[\ln T_0]$, we may also write

$$\ln T = \mu_0 - \ln g(z, \beta) + \varepsilon .$$

In particular,

(i) if $\ln T_0 \sim N(\mu, \sigma^2)$, i.e. $T_0 \sim LN(\mu, \sigma^2)$, then $\varepsilon \sim N(0, \sigma^2)$. Thus we obtain a normal regression model (*if* there is no censoring);
(ii) if $g(z, \beta) = \exp(z'\beta)$, we obtain a linear regression model: $\ln T = \mu_0 - z'\beta + \varepsilon$.

17.2.3.4 Particular Case: Weibull Baseline

In the particular case of a Weibull baseline distribution, namely $\Lambda_0(t|x) = \lambda t^\tau$, where $\alpha = (\lambda, \tau)$, along with a log-linear effect of the exogenous variable, namely $g(z, \beta) = \exp(\beta' z)$, we obtain:

$$\Lambda_{PH}(t|z, \theta) = \exp(\beta'_{PH} z)\, \lambda t^\tau$$

$$\Lambda_{AT}(t|z, \theta) = \lambda [t \exp(\beta'_{AT} z)]^\tau$$

The two models, proportional hazards and accelerated time, become therefore identical under the reparametrization $\beta_{PH} = \tau \beta_{AT}$.

17.2.4 Aggregation and Heterogeneity

Heterogeneity is the problem created by the non-observability or the omission of relevant exogenous variables. Aggregating over heterogenous individuals may create complicated structures of the hazard function. The analytical aspect is shown, for the general case, in the next lemma. An example illustrates a simple application of this lemma. Then it is shown that aggregation destroys the exponentiality of a duration.

17.2.4.1 A Basic Lemma

Let $T \mid Z \sim F_T^Z$ and $Z \sim F_Z$, i.e.

$$\Pr(T \leq t \mid Z = z) = F_T(t \mid z) \quad \text{and} \quad \Pr(Z \leq z) = F_Z(z)$$

Then

$$\begin{aligned}
f_T(t) &= \int f_T(t \mid z)\ \mathrm{d}F_Z(z) \\
S_T(t) &= \int S_T(t \mid z)\ \mathrm{d}F_Z(z) \\
\lambda_T(t) &= \frac{f_T(t)}{S_T(t)} = \frac{\int f_T(t \mid z)\ \mathrm{d}F_Z(z)}{\int S_T(t \mid z)\ \mathrm{d}F_Z(z)} \\
&= \int \lambda_T(t \mid z)\ \frac{S_T(t \mid z)}{\int S_T(t \mid z)\ \mathrm{d}F_Z(z)} \mathrm{d}F_Z(z) \\
&= \int \lambda_T(t \mid z)\ \mathrm{d}F_Z(z \mid T \geq t)
\end{aligned}$$

■

This lemma may be interpreted as follows: aggregating over heterogenous individuals, characterized by z, produces a duration distribution for which the hazard function $\lambda_T(t)$ is a *weighted* average of the individual hazard functions $\lambda_T(t \mid z)$. This possibly complicated weighting scheme may eventually account for complex hazard functions when analyzing aggregate data. A simple example illustrates this point.

17.2.4.2 An Example

Let $Z = 0$ for individuals with a low educational level, and $Z = 1$ for individuals with a high educational level. The distribution of this variable over the whole population is defined by $\Pr(Z = z) = \theta^z (1-\theta)^{1-z}$. Moreover, we suppose that:

$$(T \mid Z = j) \sim F_T^j, \quad j = 0, 1$$

Then we can deduce

$$\begin{aligned}
f_T(t) &= \theta f_T(t \mid z = 1) + (1-\theta) f_T(t \mid z = 0) \\
S_T(t) &= \theta S_T(t \mid z = 1) + (1-\theta) S_T(t \mid z = 0) \\
\lambda_T(t) &= \frac{f_T(t)}{S_T(t)} = \theta \frac{f_T^1(t)}{\theta S_T^1(t) + (1-\theta) S_T^0(t)} \\
&\quad + (1-\theta) \frac{f_T^0(t)}{\theta S_T^1(t) + (1-\theta) S_T^0(t)} \\
&= \lambda_T^1(t) \frac{\theta S_T^1(t)}{\theta S_T^1(t) + (1-\theta) S_T^0(t)} + \lambda_T^0(t) \frac{(1-\theta) S_T^0(t)}{\theta S_T^1(t) + (1-\theta) S_T^0(t)}
\end{aligned}$$

17.2.4.3 The "Mover–Stayer" Lemma

Lemma 17.1. *If* $(T \mid Z) \sim \exp\{\lambda_0(Z)\}$ *and* $Z \sim F_Z$ *arbitrary, then* $\lambda_T(t)$ *is monotone decreasing.* ■

Proof. Indeed, we successively obtain:

$$S_T(t) = \int_0^\infty S_T(t \mid z)\ \mathrm{d}F_Z(z) = \int_0^\infty \exp[-t\ \lambda_0(z)]\ dF_Z(z)$$

$$f_T(t) = -\frac{d}{\mathrm{d}t} S_T(t) = \int_0^\infty \lambda_0(z) \exp[-t\ \lambda_0(z)]\ \mathrm{d}F_Z(z)$$

$$\lambda_T(t) = \frac{f_T(t)}{S_T(t)} = \frac{\int_0^\infty \lambda_0(z) \exp[-t\ \lambda_0(z)]\ \mathrm{d}F_Z(z)}{\int_0^\infty \exp[-t\ \lambda_0(z)]\ \mathrm{d}F_Z(z)}$$

It is then easy to check that

$$\frac{d}{\mathrm{d}t}\lambda_T(t) < 0\ \forall t, \forall F_Z(Z),\ \forall \lambda_0(Z)$$

(see, for example, Fourgeaud, Gouriéroux and Pradel (1990)).

This lemma may be interpreted as follows. Individuals are characterized by their value of z. Large values of $\lambda_0(z)$ represent so-called "movers": they will leave first, while individuals represented by small value of $\lambda_0(z)$, the so-called "stayers", will leave (in probability) later. This explains why $\lambda_T(t)$ will be decreasing because being determined at each t by the remaining individuals with smaller values of $\lambda_0(z)$. This lemma also shows that although each individual duration has exponential duration, the appropriate distribution not only is not exponential but has necessarily a decreasing hazard rate, whatever is the distribution of Z.

17.2.5 Endogeneity

In the previous section, we have considered models where the covariates are exogenous. In many cases, this assumption is not realistic. Consider, for example, a model constructed in the following way: T is a duration generated conditionally on $Z = (Z_1, Z_2)$, where Z_2 is an individual characteristic and Z_1 is the level of a treatment. The variable Z_2 is known by persons who assign the treatment but unknown by the statistician. If the parameters of interest are the parameters of the conditional distribution of T given (Z_1, Z_2) these parameters are in general not identified by the conditional distribution of T given Z_1 (after integration of Z_2). Using econometric terminology, Z_1 becomes an endogenous variable. Endogeneity of treatments in duration models has been studied by Abbring and Van den Berg (2003).

17.3 Competing Risks and Multivariate Duration Models

17.3.1 Multivariate Durations

17.3.1.1 Introduction

Multivariate durations distributions are used in different situations. The first context is the analysis of multivariate elementary point processes, which occurs when we observe life lengths of several individuals belonging to the same family, or unemployment spells of couples. This is also the case when, for a given individual, we define a multivariate point process corresponding, for instance, to her labor market trajectories and to her marriage/divorce history. Another use is in point processes with more than one transition, as in the analysis of biographical data on unemployment. Yet another use is in situations where the vector of durations is latent and some sampling scheme allows one to observe only a part of this vector; this is the case in competing risks models to be presented later on.

In this section we focus our attention on general issues, namely basic definitions and properties, and methods of construction. For expository purposes we limit the presentation to bivariate distributions; extensions to more than two dimensions are fairly obvious, although notations may become cumbersome.

17.3.1.2 Basic Concepts

We start with the *multivariate survivor function* defined and denoted as

$$S_{T_1,T_2}(t_1,t_2) = \Pr(T_1 \geq t_1, T_2 \geq t_2) \; .$$

In what follows we assume that S_{T_1,T_2} is twice differentiable but in the last section we show how to treat a continuous but not everywhere differentiable survivor function as well. The *multivariate density* is defined as

$$f_{T_1,T_2}(t_1,t_2) = \frac{\partial^2}{\partial t_1 \partial t_2} S_{T_1,T_2}(t_1,t_2) \; .$$

The *marginal survivor and density functions* are defined as

$$S_{T_1}(t_1) = S_{T_1,T_2}(t_1,0)$$

$$f_{T_1}(t_1) = -\frac{d}{\mathrm{d}t_1} S_{T_1}(t_1)$$

and similarly for T_2. Often we shall write, for simplicity, $S_{1,2}$, $f_{1,2}$ or $S_j (j = 1,2)$ instead of S_{T_1,T_2}, etc.

Conditional distributions occur in different contexts and should be carefully distinguished according to the relevant conditioning event. Thus we need both $S_{1|2}(t_1|T_2 = t_2)$, $f_{1|2}(t_1|T_2 = t_2)$ and $S_{1|2}(t_1|T_2 \geq t_2)$, $f_{1|2}(t_1|T_2 \geq t_2)$. They are defined and denoted as follows:

$$S^{\geq}_{1|2}(t_1|t_2) = \Pr(T_1 \geq t_1 \mid T_2 \geq t_2) = \frac{S_{1,2}(t_1,t_2)}{S_2(t_2)}$$

$$f^{\geq}_{1|2}(t_1|t_2) = -\frac{\partial}{\partial t_1} S_{1|2}(t_1 \mid T_2 \geq t_2) = \frac{-\dfrac{\partial}{\partial t_1} S_{1,2}(t_1,t_2)}{S_2(t_2)}\,.$$

Furthermore, as shown more precisely in next subsection,

$$S^{=}_{1|2}(t_1|t_2) = \Pr(T_1 \geq t_1|T_2 = t_2) = -\frac{\dfrac{\partial}{\partial t_2} S_{1,2}(t_1,t_2)}{f_2(t_2)}$$

$$f^{=}_{1|2}(t_1|t_2) = -\frac{\partial}{\partial t_1} S^{=}_{1|2}(t_1|t_2) = \frac{f_{1,2}(t_1,t_2)}{f_2(t_2)}\,.$$

To each of these univariate conditional distributions, there corresponds a unique hazard function. For instance, marginal hazard functions are defined and denoted as:

$$\lambda_j(t_j) = \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \Pr[t_j \leq T_j < t_j + \Delta \mid T_j \geq t_j]$$

$$= -\frac{\mathrm{d} \ln\ S_j(t_j)}{\mathrm{d}t_j} = \frac{f_j(t_j)}{S_j(t_j)}$$

Conditional hazard functions are respectively defined as

$$\lambda^{\geq}_{1|2}(t_1|t_2) = \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \Pr[t_1 \leq T_1 < t_1 + \Delta \mid T_1 \geq t_1, T_2 \geq t_2]$$

$$= \frac{f^{\geq}_{1|2}(t_1|t_2)}{S^{\geq}_{1|2}(t_1|t_2)} = -\frac{\partial}{\partial t_1} \ln S_{1,2}(t_1,t_2)$$

$$\lambda^{=}_{1|2}(t_1|t_2) = \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \Pr[t_1 \leq T_1 < t_1 + \Delta \mid T_1 \geq t_1, T_2 = t_2]$$

$$= \frac{f^{=}_{1|2}(t_1|t_2)}{S^{=}_{1|2}(t_1|t_2)}$$

$$= -\frac{\partial}{\partial t_1} \left[\ln \left(-\frac{\partial}{\partial t_2} S_{1,2}(t_1,t_2) \right) \right]$$

17.3.1.3 Construction of Multivariate Distributions

Several techniques for constructing multivariate distributions are worth mentioning. The most trivial one is the case of independent components in which case the joint survivor and density functions are the products of (arbitrary) corresponding marginal functions, and in which the conditional survivor, density and hazard functions coincide with the corresponding marginal functions.

For the dependent case, two general procedures are: (i) take two univariate distributions, choose one to be marginal and take the other one to be conditional to the first by making its parameters to be a function of the conditioning variable; (ii) take a joint distribution with survivor $S(t_1,t_2,y)$ where y is an auxiliary variable such that $S(t_1,t_2 \mid y)$ is meaningful, and marginalize it into $S_{1,2}(t_1,t_2)$.

17.3.2 Competing Risks Models: Definitions

Competing risks duration models may be applied to situations where the state space E has more than two elements: $E = \{E_0,E_1,\ldots,E_J\}$, $J > 2$. Such models involve specifying not only the date at which the process leaves the initial state E_0, but also which state in $\{E_1,\ldots E_J\}$ is entered.

Consider, for instance, a medical trial where a patient is submitted to a "treatment" for a supposedly known disease and where the survival time is observed. Typically, the cause of death is multiple; in particular, it may be different from the disease for which the treatment was originally designed, and the cause is possibly associated with the treatment itself. One says that several risks "compete" to cause the death of the patient. Similarly, in the labor market, when the initial state E_0 is unemployment, it may be relevant to distinguish several exit states, for example full-time employment, part-time employment or early retirement. The relevance of these distinctions is based on the fact that economic, social and institutional factors may be important to explain both durations and transitions of the individual trajectories; in other words, they are particularly important when analyzing biographical data.

Thus the data have the form (T,K) where T is the sojourn duration in the initial state and K is the destination state. Therefore the law of such a process is specified by the so-called sub-distribution

$$\Pr(T \geq t, K = k) = \Pr(T_j \geq T_k \geq t, \quad \forall j \neq k)$$

Competing risk models provide a specification of $\Pr(T \geq t, K = k)$ based on the following idea. T represents the duration of sojourn in the initial state E_0, whatever the destination state is. The latent random variable T_j would represent the duration of sojourn in the initial state if E_j were the only possible destination. In the competing risk models, if ties have zero probability, i.e. $\Pr(T_i = T_j) = 0$, $\forall i \neq j$, the T_j's are connected by the relationships:

$$T = \min_j\{T_j\}, \quad j = 1,\cdots J,$$
$$K = \ \operatorname{argmin}_j\{T_j = T\}\ .$$

Thus, the T_j's are latent duration variables because only their minimum is observed. This structure permits to write easily the marginal laws of T and K, which are given by:

$$S_T(t) = \Pr(T \geq t) = \Pr\left\{\cap_{j=1,\ldots,K}(T_j \geq t)\right\}$$
$$\Pr[K = k] = \Pr\left\{\cap_{j\neq k}(T_k < T_j)\right\}$$

Intuitively, k is the index of the lowest latent duration (given an ascending order on the $j's$). In order to evaluate the likelihood function, we start by the joint survivor function, using * as an upper index in the notation of the joint distribution of the latent durations $(T_1,\ldots,T_J)$ to stress that those durations are latent:

$$S^*(t_1,\ldots,t_J) = \Pr(T_1 \geq t_1,\ldots,T_J \geq t_J)$$

for any $(t_1,\ldots t_J) \in \mathbb{R}_+^J$. The survivor function of the observed duration $T = \min_j(T_j)$ satisfies

$$S_T(t) = S^*(t,\ldots,t),\ \ t \in \mathbb{R}^+$$

The marginal survivor function of the latent duration T_j, for $j = 1,\ldots,J$, is denoted ${S_j}^*$ and defined as:

$${S_j}^*(t_j) = S^*(0,\ldots,0,t_j,0,\ldots,0)\ .$$

In the case where the T_j's are independent, we have

$$S^*(t_1,\ldots,t_J) = \Pi_{j=1}^J {S_j}^*(t_j)\ .$$

Now, let us suppose that the functions S^* and consequently S_T and ${S_j}^*$ are continuously differentiable. The marginal and relevant conditional hazard functions of the latent duration T_j, for $j = 1,\ldots,J$, are denoted and defined as

$$\lambda_j(t) = \lim_{\Delta\downarrow 0}\frac{1}{\Delta}\Pr(t \leq T_j < t+\Delta \mid T_j \geq t) = -\mathrm{d}\ln {S_j}^*(t)/\mathrm{d}t, \quad t \in \mathbb{R}_+,$$
$$\lambda_{j|T}^{\geq}(t) = \lim_{\Delta\downarrow 0}\frac{1}{\Delta}\Pr(t \leq T_j < t+\Delta \mid T \geq t)$$
$$= -\frac{\partial}{\partial t_j}\ln S^*(t_1,\ldots,t_J)\mid_{t_1=t_2=\cdots=t_J=t}$$

where $\lambda_{j|T}^{\geq}(t)$ is a short cut for $\lambda_{T_j|T}^{\geq}(t|t)$. When the $T_j's$ are mutually independent, it is obvious that:

$$\lambda_{j|T}^{\geq}(t) = \lambda_j(t), \qquad \text{for any } t \in \mathbb{R}_+\ .$$

The hazard function of the observed duration T is denoted and defined as

$$\begin{aligned}\lambda_T(t) &= \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \Pr(t \le T < t+\Delta \mid T \ge t) \\ &= -\mathrm{d}\ln S_T(t)/dt\,,\ t \in_+ . \\ &= \sum_{j=1}^{J} \lambda^{\geq}_{j|T}(t)\end{aligned}$$

because, in the definition of $h_T(t)$, the derivative of $S_T(t)$ is a directional derivative (in the direction of the main diagonal $(1,1,\ldots,1)$) of $S^*(t_1,\ldots,t_J)$. In the continuously differentiable case, the likelihood function may be evaluated by differentiating the sub-distribution, namely:

$$\begin{aligned}l_{T,K}(t,k) &= -\frac{d}{\mathrm{d}t}\Pr(T \ge t, K = k) \\ &= -\frac{d}{\mathrm{d}t}\Pr\left\{\cap_{j\neq k}(T_j > T_k \ge t)\right\}\end{aligned}$$

Remember that a basic result of differential calculus gives:

$$S^*(t_1,\ldots,t_J) = -\int_{u=t_k}^{\infty} \frac{\partial}{\partial t_k} S^*(t_1,\ldots,t_j)\,\mathrm{d}u$$

and, similarly, a basic result of conditional probability gives:

$$S^*(t_1,\ldots,t_J) = \int_{u=t_k}^{\infty} S^{=}_{\bar{k}|k} = (t_1,\ldots,t_{k-1},t_{k+1},\ldots,t_J \mid T_k = u)\, f_k(u)\,\mathrm{d}u$$

where $\bar{k} = \{1,2,\ldots,j\} \setminus \{k\}$ and

$$\begin{aligned}&S^{=}_{\bar{k}|k}(t_1,\ldots,t_{k-1},t_{k+1},\ldots,t_J \mid T_k = t_k) \\ &= \Pr(T_1 \ge t_1,\ldots,T_{k-1} \ge t_{k-1}, T_{k+1} \ge t_{k+1},\ldots,T_J \ge t_J \mid T_k = t_k)\end{aligned}$$

Thus the likelihood function may be written as:

$$S^{=}_{\bar{j}|j}(t_1,\ldots,t_{j-1},t_{j+1},\ldots,t_J \mid t_j) = -\frac{\frac{\partial}{\partial t_j} S^*(t_1,\ldots,t_J)}{f_j(t_j)}\,.$$

In the sequel we use the following simplified notation

$$S^{=}_{\bar{j}|j}(t) = S^{=}_{\bar{j}|j}(t,t,\ldots,t \mid t)$$

Then, the sub-distribution may be written as:

$$\int_{t_k}^{\infty} S_{\bar{k}|k}(u) f_k(u)\mathrm{d}u$$

Therefore,

$$\begin{aligned} l_{T,K}(t,k) &= -\frac{d}{dt}\int_t^\infty S^{=}_{\bar{k}|k}(u) f_k(u)\mathrm{d}u \\ &= S^{=}_{\bar{k}|k}(t) \times f_k(t) \\ &= -\frac{\partial}{\partial t_k} S^*(t,\ldots,t) \\ &= -S^*(t,\ldots t) \times \frac{\partial}{\partial t_k} \ln S^*(t,\ldots,t) \end{aligned}$$

Using a disjunctive coding for the exit state, namely

$$A = (A_1,\ldots,A_J), \quad A_j = \mathbb{I}\{K=j\}$$

we may also write

$$l_{T,A}(t,a) = \prod_{j=1}^{J}\left[f_j(t)S^{=}_{\bar{j}|j}(t)\right]^{a_j} = S_T(t)\prod_{j=1}^{J}\left[\lambda^{\geq}_{j|T}(t)\right]^{a_j} .$$

In case of independent latent durations, we have:

$$\begin{aligned} l_{T,K}(t,k) &= f_k(t)\prod_{j\neq k} S_j(t) \\ &= \lambda_k(t) S_T(t) . \end{aligned}$$

17.3.3 Identifiability of Competing Risks Models

The basic idea of competing risks models is to interpret the data (T,K), representing the sojourn duration in the initial state and the label of the exit state, as the observation of the minimum component of a random vector along with the coordinate where the minimum is obtained. Intuition suggests that these observations give no information on the question whether the coordinate of the random vector, i.e. of the latent durations, are independent or not. This intuition is confirmed by next theorem

Theorem 17.1. *Let us denote $\mathscr{S} = \{S^*(t_1\cdots t_J)\}$ the set of J-dimensional survivor functions, $\mathscr{S}_I = \{S^* \in \mathscr{S} \mid S^*(t_1\cdots t_J) = \prod_j S_j(t_j)\}$ the subset of J-dimensional survivor functions with independent components, $l^*(t,k)$ the likelihood function for a model in $\mathscr{S}$, and $l_I(t,k)$ the likelihood function for a model in $\mathscr{S}_I$. Then:*

$$\forall S^* \in \mathscr{S}, \quad \exists!\, S_I \in \mathscr{S}_I \text{ such that } l^*(t,k) = l_I(t,k)$$

In particular,

$$\lambda_{j|T}^{*,\geq}(t) = \lambda_{j,I}(t)$$

■

In the continuous case, the proof of this theorem comes from the fact that, in the general case, $l^*(t,k) = \lambda_{k|(t)}^{*,\geq} S_T(t)$ and that $\lambda_T(t) = \sum_j \lambda_{j|T}^{*,\geq}(t)$, i.e. the distribution of the observed duration depends only on the sum of the conditional hazard functions. Therefore the equality $\lambda_{k|T}^{*,\geq}(t) = \lambda_{k,I}(t)$ ensures the equality of likelihood functions. Mouchart and Rolin (2002) gives a slightly more general statement and proof of this theorem.

This theorem means that to any competing risks model with dependent latent durations, one may associate an observationally equivalent model with independent latent durations. The rule of association is simply to build the joint latent distribution with marginal hazard functions of the independent model that are equal to the conditional hazard functions of the dependent model. To illustrate this point, we can consider the following bivariate example. Suppose that the joint survivor function of the two latent durations (T_1, T_2) is given by:

$$S^*(t_1,t_2) = \exp\Big\{1 - \alpha_1 t_1 - \alpha_2 t_2 - \exp\left[\alpha_{12}(\alpha_1 t_1 + \alpha_2 t_2)\right]\Big\}$$

where $\alpha_1, \alpha_2 > 0$ and $\alpha_{12} > -1$. Here the parameter α_{12} measures the dependence between the two latent durations T_1 and T_2 in the sense that T_1 and T_2 are independent once $\alpha_{12} = 0$. The conditional and marginal hazard functions of this model are respectively:

$$\lambda_{j|T}^{*,\geq}(t) = \alpha_j\Big\{1 + \alpha_{12}\exp[\alpha_{12}(\alpha_1+\alpha_2)t]\Big\}, \quad j = 1,2$$

and

$$\lambda_j^*(t) = \alpha_j[1 + \alpha_{12}\exp(\alpha_j\alpha_{12}t)], \quad j = 1,2\,.$$

Marginal survivor functions are then

$$S_j^*(t_j) = \exp\Big[1 - \alpha_j t_j - \exp(\alpha_{12}\alpha_j t_j)\Big], \quad j = 1,2\,,$$

from which it is obvious that

$$S^*(t_1,t_2) \neq S_1^*(t_1)S_2^*(t_2), \quad (t_1,t_2) \in \mathbb{R}_+^2$$

except if $\alpha_{12} = 0$. The likelihood element of an observation (t,k) may be written as

$$\begin{aligned} l^*(t,k) = {} & \alpha_k\{1 + \alpha_{12}\exp[\alpha_{12}(\alpha_1+\alpha_2)t]\} \\ & \times \exp\{1 - (\alpha_1+\alpha_2)t - \exp[\alpha_{12}(\alpha_1+\alpha_2)t]\} \end{aligned}$$

The observationally equivalent model (i.e. having the same likelihood function) with independent latent durations has marginal hazard functions given by $\lambda_{j|T}^{*,\geq}(t)$ above and eventually marginal and joint survivor functions of latent durations given by:

$$S_{j,I}(t_j) = \exp\left\{ \frac{\alpha_j}{\alpha_1+\alpha_2} - \alpha_j t_j - \frac{\alpha_j}{\alpha_1+\alpha_2} \exp\alpha_{12}(\alpha_1+\alpha_2)t_j \right\}, \quad j=1,2$$

$$S_I(t_1,t_2) = \exp\left\{ 1-\alpha_1 t_1 - \alpha_2 t_2 - \frac{1}{\alpha_1+\alpha_2}\left[\alpha_1 \exp\alpha_{12}(\alpha_1+\alpha_2)t_1 \right.\right.$$

$$\left.\left. +\alpha_2 \exp\alpha_{12}(\alpha_1+\alpha_2)t_2\right] \right\}$$

Note that the latent models are clearly different unless $\alpha_{12}=0$, i.e.

$$S^*(t_1,t_2) \neq S_I(t_1,t_2)$$

but the statistical models are observationally equivalent, i.e. $l^*(t,k) = l_I(t,k)$. Note also that both latent models have been identifiably parametrized, but the parameters have very different meaning in the two latent models. In particular, α_{12} measures the association among the latent variables in the case of dependence whereas α_{12} is a common parameter of the two marginal distributions in the case of independent latent variables. The identifiability of the competing-risks duration model with unobserved heterogeneity has been studied by Heckman and Honoré (1989). Their results have been completed by those obtained by Honoré (1993) for duration models with multiple spells and with unobserved heterogeneity.

17.3.4 Right-Censoring

One usual feature of duration data is that the sampling scheme often produces right-censored observations, i.e. observations which have not yet left the initial state E_0 at the end of the sampling period. For example, in the case of single-spell unemployment duration data, the sampling scheme is often the following. Individual observations are sampled from the inflow of individuals entering unemployment at time t_0 and followed up until date C, which is possibly determined by the researcher. Now let us assume that C is greater than t_0. Some observations correspond to individuals leaving the unemployment status before C, in which case they generate complete unemployment durations. Other sampled individuals have not left the unemployment state at date C and so they generate right-censored unemployment durations. Rather than sampling from the inflow into unemployment at a given date t_0, the analyst may sample from inflows considered at several staggered dates $t_0^1, t_0^2, \ldots$ and follow up observations once again up to a censoring time C. Right-censoring can be modelled using the framework of the competing risks models with state space $\{E_0, E_1, \ldots, E_J\}, J>1$, where the last state E_J denotes the right-censored situation. To illustrate this kind of formalization, let us consider a bivariate competing risks model (T_1, T_2) with state space $\{E_0, E_1, E_2\}$, E_0 labelling unemployment, E_1 employment and E_2 right-censoring. Thus $T_2 = C$. In other words, censoring is often associated with a residual state in a model with multiple states. Suppose first that all individual observations are sampled at the same date t_0. Without loss of generality,

one may write $t_0 = 0$ (after some relevant time translation). Within the framework presented in the previous section, this model may be viewed as resulting from a latent survivor function $S_{1,2}(t_1, t_2 \mid \theta)$ with parameter θ, and a generic element of the likelihood function may be written as:

$$l_{T,D}(t,d) = \left[f_1(t \mid \theta) S^{=}_{2|1}(t \mid \theta)\right]^d \left[f_2(t \mid \theta) S^{=}_{1|2}(t \mid \theta)\right]^{1-d}$$

where $D = \mathbb{I}_{\{T_1 \leq T_2\}}$. In view of the identification problem, and because in many cases censoring mechanisms are independent of the unemployment process, it is often assumed that T_1 and T_2 are independent. Then,

$$l_{T,D}(t,d) = [f_1(t \mid \theta) S_2(t \mid \theta)]^d [f_2(t \mid \theta) S_1(t \mid \theta)]^{1-d}$$

If moreover θ may be factorized into $\theta = (\theta_1, \theta_2)$, such that θ_1 characterizes the distribution of T_1 and θ_2 the distribution of T_2, the likelihood reduces to

$$l_{T,D}(t,a) = L_1(\theta_1) L_2(\theta_2)$$

where

$$L_1(\theta_1) = f_1(t \mid \theta_1)^d S_1(t \mid \theta_1)^{1-d} .$$

The parameters of interest are in general those of the distribution of duration T_1, and their estimation could be deduced from $L_1(\theta_1)$ only. Then the generic element of the relevant factor of the likelihood function is $f_1(t \mid \theta_1)$ (resp. $S_1(t \mid \theta_1)$) for an uncensored (resp. a right-censored) observation.

Another model generating censored data may be the following one. Let T_0 be the age of an individual entering unemployment. This age is randomly generated by the individual previous labor market history. The duration of the unemployment spell is T_1 and the age at the end of the unemployment spell is then $T_0 + T_1$. The econometric model specifies the joint distribution of (T_0, T_1) and these two random variables are not, in general, assumed to be independent. A natural specification could be a sequential one: the (marginal) distribution of T_0 is first specified and a conditional distribution of T_1 given T_0 completes the model.

Let us now assume that all the individuals are observed at a given date T_*. In general this date is also random but, for simplicity, we consider T_* as fixed (the model is conditional to T_*). Let us also assume that the sample is constructed in such a way that $T_0 \leq T_*$ (all the individuals have entered unemployment). Then T_0 is always observed but T_1 is not censored if $T_0 + T_1 \leq T_*$. Otherwise, the unemployment spell duration is censored.

Let us define $T_2 = T_* - T_0$. From the distribution of (T_0, T_1) we obtain the distribution of (T_1, T_2), and we may consider the observations as generated by a censored duration model: T_1 is observed only if $T_1 \leq T_2$. But the following specification of a likelihood based on the generic element:

$$l_{T,D}(t,d) = f_1(t)^d S_1(t)^{1-d}$$

where $T = \min(T_1, T_2)$, $D = \mathbb{I}(T_1 \leq T_2)$, f_1 and S_1 are the density and the survivor functions of T_1, is incorrect for two reasons:

(i) First if T_0 and T_1 are dependent, T_1 and T_2 are also dependent and the likelihood function must be based on their joint distribution.
(ii) The censoring mechanism is different from the usual competing risks model because T_0 or T_2 is always observed and the likelihood of the actual data must be the density of (T_2, T, D). The generic element of this likelihood is then

$$l_{T_2,T,D}(t_2, t, d) = f_2(t_2)\, f^{=}_{1|2}(t \mid t_2)^a\, S^{=}_{1|2}(t \mid t_2)^{1-d}$$

using our previous notations. Finally, note that the identification result of Sect. 17.3.3 does not apply to this case since the censoring mechanism is different from the competing risks model.

17.4 Inference in Duration Models

17.4.1 Introduction

Models actually used in econometrics for dealing with duration data are characterized by two noteworthy features: durations are non-negative random variables and most data sets involve right-censored data. In this section, we focus our attention on the implications of censoring, both for adapting the inference procedure and for evaluating the consequences of misspecification. We first review the inference in parametric models, both in the marginal and in the conditional case, with a particular attention on a rigorous specification of the likelihood function; next we consider non- and semi-parametric models. In each case, we first specify the structure of the model and next give some illustrations with significantly relevant particular cases.

17.4.2 Parametric Models

17.4.2.1 Inference in Marginal Models

The Basic Model

The basic model considers a random censoring process that is independent of the duration variable. Let us introduce the following notations:

$$
\begin{aligned}
\eta &= (\eta_1,\ldots,\eta_n)' \quad \text{denote } \textit{latent durations},\\
\zeta &= (\zeta_1,\ldots,\zeta_n)' \quad \text{denote } \textit{latent censoring indicators},\\
T &= (T_1,\ldots,T_n)', \text{ with } T_i=\eta_i\wedge\zeta_i, \text{ are } \textit{observed durations}\\
D &= (D_1,\ldots,D_n)', \text{ with } D_i=\mathbb{I}_{\{\eta_i\le\zeta_i\}}=\mathbb{I}_{\{T_i=\eta_i\}},\\
X &= (X_1,\ldots,X_n)', \text{ with } X_i=(T_i,D_i)', \text{ denote } \textit{complete data}\\
X &= (T,D) \text{ with } \dim(X)=(n\times 2,1)\\
&\phi \text{ is a sufficient parametrisation for the process generating } (\eta,\zeta)
\end{aligned}
$$

Assumptions

A.1 (independent sampling): $\perp\!\!\!\perp_i(\eta_i,\zeta_i)\mid\phi$
A.2 (independent censoring): $\eta_i\perp\!\!\!\perp\zeta_i\mid\phi$
A.3 (definition of θ as a sufficient parametrization for η): $\eta_i\perp\!\!\!\perp\phi\mid\theta$
A.4 (definition of ω as a sufficient parametrization for ζ): $\zeta_i\perp\!\!\!\perp\phi\mid\omega$
A.5 (variation-free parameters) : $(\theta,\omega)\in\Theta_\theta\times\Theta_\omega$
A.6 θ is the only parameter of interest

Latent Likelihood

Under (A.1) to (A.5), the complete latent likelihood is therefore:

$$L^{**}(\phi)=\prod_i f_\eta(\eta_i\mid\theta)\cdot\prod_i f_\zeta(\zeta_i\mid\omega)=L_1^*(\theta)\cdot L_2^*(\omega)$$

Under (A.6), the relevant latent likelihood is

$$L_1^*(\theta)=\prod_i f_\eta(\eta_i\mid\theta)=f_\eta(\eta\mid\theta)$$

Actual Likelihood

Considering the actually available data, namely (T,D), the complete actual likelihood is

$$
\begin{aligned}
L(\phi) &= \prod_i f_\eta(T_i\mid\theta)^{D_i}S_\eta(T_i\mid\theta)^{1-D_i}\prod_i f_\zeta(T_i\mid\omega)^{1-D_i}S_\zeta(T_i\mid\omega)^{D_i}\\
&= L_1(\theta)L_2(\omega)
\end{aligned}
$$

Under (A.6), the relevant actual likelihood is:

$$L_1(\theta)=\prod_i f_\eta(T_i\mid\theta)^{D_i}S_\eta(T_i\mid\theta)^{1-D_i}=\prod_i\lambda_\eta(T_i\mid\theta)^{D_i}S_\eta(T_i\mid\theta)$$

Thus the logarithm of the relevant actual likelihood is:

$$L(\theta) = \ln L_1(\theta) = \sum_i D_i \ln f_\eta(T_i \mid \theta) + \sum_i (1 - D_i) \ln S_\eta(T_i \mid \theta)$$

$$= \sum_i [D_i \ln \lambda_\eta(T_i \mid \theta) + \ln S_\eta(T_i \mid \theta)]$$

$$= \sum_i [D_i \ln \lambda_\eta(T_i \mid \theta) - \Lambda_\eta(T_i \mid \theta)]$$

The Exponential Case

The consequences of censoring are best understood by considering with some detail the case where the duration of interest is exponentially distributed, which means that $f_\eta(\eta_i \mid \theta) = \theta e^{-\theta \eta_i}$ while $f_\zeta(\zeta_i \mid \omega)$ is left unspecified. Thus, the latent process generating η is a member of the exponential family, $\sum_i \eta_i = \eta_+$ is a minimal sufficient complete statistic of the latent process and, for a sample of size n, the Fisher information is $n\theta^{-2}$. With censoring, the relevant actual likelihood is written as:

$$L(\theta) = \sum_i D_i \ln \theta - \sum_i T_i \theta = (\ln \theta) D_+ - \theta T_+$$

where $D_+ = \sum_i D_i$ and $T_+ = \sum_i T_i$. The score and the statistical information are accordingly:

$$S(\theta) = \frac{d}{\mathrm{d}\theta} L(\theta) = \frac{D_+}{\theta} - T_+$$

$$J(\theta) = -\frac{d^2}{\mathrm{d}\theta^2} L(\theta) = \frac{D_+}{\theta^2}$$

taking into account that $J(\theta)$ and therefore $I(\theta)$ are block diagonal. Therefore the maximum likelihood estimator of θ is:

$$\hat{\theta}_{ML} = \frac{D_+}{T_+}$$

Let us recall that:

$$\sqrt{n}(\hat{\theta}_{ML,n} - \theta) \xrightarrow{\mathscr{L}} N\left\{0, [I(\theta)]^{-1}\right\}$$

where

$$I(\theta) = V\left\{\frac{d}{\mathrm{d}\theta} L(\theta) \mid \theta\right\} = \mathbb{E}[J(\theta) | \theta] = \frac{E[D_+ \mid \theta]}{\theta^2}$$

Note that:

$$E[D_i \mid \phi] = \Pr[\eta_i \leq \zeta_i \mid \phi] = E[F_\eta(\zeta_i \mid \theta) \mid \phi] = 1 - E[e^{-\theta \zeta_i} \mid \phi]$$

Therefore:

$$E[D_i \mid \theta] = 1 - E[e^{-\theta \zeta_i} | \theta] = 1 - \int e^{-\theta \zeta_i} \mathrm{d}F_\zeta(\zeta_i)$$

In practice, $I(\theta)$ is estimated as:

$$I(\hat{\theta}_{MV,n}) = \frac{D_+}{\hat{\theta}^2_{ML,n}}$$

Let us turn now to the uncensored case. In the model with censoring, there is only one parameter, $\theta \in \mathbb{R}_+$, and the bivariate statistic (D_+, T_+) is minimal sufficient but not complete. This is an example of a curved exponential family with canonical parameter $(\theta, \ln \theta)$. Also, let us notice the differences in the maximum likelihood estimations:

$$D_+ \longrightarrow n > D_+$$

$$L(\theta) = (\ln \theta) D_+ - \theta T_+ \longrightarrow n \ln \theta - \theta T_+$$

$$\hat{\theta}^c_{ML} = \tfrac{D_+}{T_+} \longrightarrow \hat{\theta}^{nc}_{ML} = \tfrac{n}{T_+} > \tfrac{D_+}{T_+}$$

In other words, the cost of overlooking censoring may be appreciated by considering the difference between the (true) Fisher information, and the numerical value of the maximum likelihood estimator:

$$\frac{\hat{\theta}^c_{ML}}{\hat{\theta}^{nc}_{ML}} = \frac{D_+}{n} \leq 1 \quad \text{and} \ = 1 \Longleftrightarrow D_+ = n$$

17.4.2.2 Inference in Conditional Models

The General Statistical Model

Let us introduce the following definitions and assumptions:

- $\theta = (\alpha, \beta) \in \Theta_\alpha \times \Theta_\beta \subset \mathbb{R}^{k_\alpha} \times \mathbb{R}^{k_\beta}$, $\quad k_\alpha$ and k_β finite.
- *Data*:

$$Y_i = (T_i, D_i),\ Y = (Y_1, \ldots, Y_n)$$

$$X_i = (Y_i, Z_i),\ X = (X_1, \ldots, X_n)$$

- *Definition of* κ *and* θ : $Z \perp\!\!\!\perp \theta \mid \kappa$ and $Y \perp\!\!\!\perp \kappa \mid Z, \theta$
- *Assumptions*
  variation-free parameters: $(\kappa, \theta) \in \Theta_\kappa \times \Theta_\theta$
  conditional independence: $\perp\!\!\!\perp_i Y_i \mid Z, \theta$ and $Y_i \perp\!\!\!\perp Z \mid Z_i, \theta$
  θ is the only parameter of interest

Therefore, the relevant actual loglikelihood takes the form:

$$L(\theta) = \sum_i D_i \ln f_\eta(T_i \mid z_i, \theta) + \sum_i (1 - D_i) \ln S_\eta(T_i \mid z_i, \theta)$$

$$= \sum_i D_i \ln l_\eta(T_i \mid z_i, \theta) - \sum_i \Lambda_\eta(T_i \mid z_i, \theta)$$

The score and the statistical information are equal to:

$$S(\theta) = \frac{d}{\mathrm{d}\theta} L(\theta) = \sum_i \frac{D_i}{\lambda_\eta(T_i \mid z_i, \theta)} \frac{d}{\mathrm{d}\theta} \lambda_\eta(T_i \mid z_i, \theta) - \sum_i \frac{d}{\mathrm{d}\theta} \Lambda_\eta(T_i \mid z_i, \theta)$$

$$I(\theta) = \frac{-d^2}{\mathrm{d}\theta \mathrm{d}\theta'} L(\theta) = \sum_i D_i [\lambda_\eta(T_i|z_i, \theta)]^{-2} \frac{d}{\mathrm{d}\theta} \lambda_\eta(T_i|z_i, \theta) \frac{d}{\mathrm{d}\theta'} \lambda_\eta(T_i|z_i, \theta)$$
$$- \sum_i D_i \lambda_\eta(T_i|z_i, \theta)^{-1} \frac{d^2}{\mathrm{d}\theta \mathrm{d}\theta'} \lambda_\eta(T_i|z_i, \theta) + \sum_i \frac{d^2}{\mathrm{d}\theta \mathrm{d}\theta'} \Lambda_\eta(T_i|z_i, \theta)$$

Notice once more that the expectation of $I(\theta)$ depends both on θ and K, and thus, on the parameter of the censoring variable.

The Proportional Hazard Model

When

$$\lambda_\eta(t \mid z, \theta) = g(z, \beta) \lambda_0(t \mid \alpha) \,,$$

the log-likelihood function may be written as:

$$L(\theta) = \sum_i D_i \ln \ \lambda_\eta(T_i \mid z, \theta) - \sum_i \Lambda_\eta(T_i \mid z, \beta)$$
$$= \sum_i D_i \ln \ g(z_i, \beta) + \sum_i D_i \ \ln \ \lambda_0(T_i|\alpha) - \sum_i g(z_i, \beta) \Lambda_0(T_i|\alpha)$$

and, under the log-linear specification $g(z, \beta) = \exp\left(z^{'} \beta\right)$:

$$L(\theta) = \beta' \sum_i D_i z_i + \sum_i D_i \ln \lambda_0(T_i|\alpha) - \sum_i e^{z_i' \beta} \Lambda_0(T_i|\alpha)$$

The Mixed Proportional Hazard Model and its Identifiability

The mixed proportional hazard (MPH) model is characterized by the following hazard function:

$$\lambda_T(t \mid z) = \lambda_0(t) \, g(z) \, \upsilon$$

where $\lambda_0(t)$ is a baseline hazard function, $g(z)$ is the function measuring the proportional effect of observable covariates z on the hazard function, and υ is an individual-specific random term representing unobserved individual heterogeneity. The cumulative density function of υ is denoted H. This model is supposed to verify the following assumptions:

Assumption 1: The covariate vector z is a finite-dimensional vector of dimension k $(1 \leq k \leq \infty)$. The function $g(z)$ is positive for every $z \in \mathscr{Z} \subset \mathbb{R}^k$.

Assumption 2: The function $\lambda_0(t)$ is positive and continuous on $[0,\infty)$, except that $\lim_{t\to 0}\lambda_0(t)$ may be infinite. For every $t \geq 0$,

$$\int_0^t \lambda_0(u)\mathrm{d}u < \infty \quad \text{while} \quad \lim_{t\to\infty}\int_0^t \lambda_0(u)\mathrm{d}u = \infty$$

Assumption 3: The distribution H of the random term υ in the inflow (i.e. when $t = 0$) satisfies $\Pr\{\upsilon \in]0,\infty)\} = 1$.

Assumption 4: The individual value of υ is time-invariant.

Assumption 5: In the inflow (i.e. when $t = 0$), υ is independent of z.

This model is nonparametrically identified if there is a unique set of functions λ_0, g and H that generates the observable distribution of the data, namely $F(t \mid z)$. Conditions for identification are the following (see Van den Berg (2001), for a clear exposition):

Assumption 6: (variation in observed covariates): The set $\mathscr{Z}$ of possible values of z contains at least two values, and $g(z)$ is not constant on $\mathscr{Z}$.

Assumption 7: (normalizations): For some a priori chosen t_0 and z_0, there holds:

$$\int_0^{t_0} \lambda_0(u)\,\mathrm{d}u = 1 \quad \text{and} \quad g(z_0) = 1$$

Assumption 8: (tail of the unobserved heterogeneity distribution): $E(\upsilon) < \infty$.

Assumptions 6 and 8 can be alternatively stated:

Assumption 6b: (variation in observed covariates): The vector z includes an element z^a such that the set $\mathscr{Z}^a$ of its possible values contains a non-empty open interval. For given values of the other elements of z, the value of z^a varies over this interval. Moreover, $g(z)$ as a function of z^a is differentiable and not constant on this interval.

Assumption 8b: (tail of the unobserved heterogeneity distribution): The random variable υ is continuous, and the probability density function $h(\upsilon)$ of υ verifies the following property:

$$\lim_{\upsilon\to\infty} \frac{h(\upsilon)}{\upsilon^{-1-\in}V(\upsilon)} = 1$$

where $\varepsilon \in [0,1]$ is fixed in advance, and where $V(\upsilon)$ is a function such that:

$$\lim_{s\to\infty} \frac{V(s\upsilon)}{V(s)} = 1\,.$$

Identification of the MPH model has been analyzed successively by Elbers and Ridder (1982), Heckman and Singer (1984b), Ridder (1990), Melino and Sueyoshi (1990), and Kortram, Lenstra, Ridder and Van Rooij (1995).

The Accelerated Life Model

When $\lambda_\eta(t \mid z,\theta) = g(z,\beta)\ \lambda_0(t\, g(z,\beta) \mid \alpha)$, the log-likelihood function, for an arbitrary family of baseline distributions, may be written as:

$$L(\theta) = \sum_i D_i[\ln g(z_i,\beta) + \ln \lambda_0(T_i\, g(z_i,\beta)|\alpha)] - \sum_i \Lambda(T_i\, g(z_i,\beta)|\alpha)$$

When the baseline distribution is exponential, namely when $\lambda_0(t_i|\alpha) = \alpha$, we obtain:

$$L(\theta) = \sum_i D_i[\ln g(z_i,\beta) + \ln \alpha] - \alpha \sum_i T_i g(z_i,\beta)$$

In the particular case where $g(z_i,\beta) = \exp\left(z_i'\beta\right)$, we obtain a proportional hazard model. More generally, this is also the case for a Weibull baseline distribution:

$$L(\theta) = \ln \alpha \sum_i D_i + \beta' \sum_i D_i z_i - \alpha \sum_i T_i e^{z_i'\beta}$$

17.4.3 Non-parametric and Semi-parametric Models

17.4.3.1 Marginal Models: The Kaplan–Meier Estimator of the Survivor Factor

If we want to estimate $S_T(t)$ in presence of right-censoring, a simple idea is to adjust the hazard rates of the product form of the (discrete) empirical survivor function. With the same data as for the parametric models:

$$\begin{aligned} Y_i &= (T_i, D_i) \\ T_i &= \min(\eta_i, \zeta_i) \\ D_i &= \mathbb{I}_{\{T_i=\eta_i\}} \end{aligned}$$

we now evaluate:

$$\begin{aligned} T_i &\to T_{(1)} < T_{(2)}, \ldots, T_{(n)} \quad \text{(order statistics)} \\ D_i &\to D_1', D_2', \ldots, D_n' \text{: (censoring indicators corresponding to the } T_{(i)}) \\ R(t) &= \sum_i \mathbb{I}_{\{T_{(i)} \geq t\}} \\ B(T_{(i)}) &= \sum_j D_j' \mathbb{I}_{\{T_j = T_{(i)}\}} \end{aligned}$$

Thus $R(t)$ represents the number of individuals at risk at time t, i.e. those who are neither "dead" nor censored at time t^-, and $B(T_{(i)})$ represents the number of deaths (i.e. exiting without being censored) at the observed time $T_{(i)}$. A natural way of taking censoring into account is to consider that at the time $T_{(i)}$, $B(T_{(i)})$ is the realization of a binomial variable with parameter $\left(R\left(T_{(i)}\right), \lambda\left(T_{(i)}\right)\right)$. Then the hazard function at (observed) time $T_{(i)}$ and the survivor functions are estimated as:

$$\hat{\lambda}(T_{(i)}) = \frac{B(T_{(i)})}{R(T_{(i)})}$$

$$\hat{S}_{KM}(t) = \prod_{\{T_{(i)}<t\}} [1-\hat{\lambda}(T_{(i)})]$$

Remarks

1. If at $T_{(i)}$ there are only censored data, we have $B(T_{(i)}) = 0$ and therefore $\hat{S}_{KM}(T_{(i)})$ is continuous at $T_{(i)}$.
2. If the largest observation is a censored one, $\hat{S}_{KM}(t)$ is strictly positive and continuous, at $T_{(n)}$:

$$\hat{S}_{KM}(t) = \hat{S}_{KM}(T_{(n)}) > 0, \qquad \forall t > T_{(n)}$$

   If furthermore $T_{(n-1)}$ is not censored, $\lim_{t\to\infty} \widehat{F}_{KM}(t) > 0$, which means that F could be defective. A natural interpretation of this occurrence, in the case of a life duration, is the following: if the largest observation does not correspond to an exit (or a death), there is no empirical reason not to believe that such a life could possibly be infinite. If one is willing to avoid defective distributions, one may modify the Kaplan–Meier estimator as follows:

$$\widehat{S}^m_{KM}(t) = \prod_{\{T_{(i)}\le t\}} [1-\hat{h}(T_{(i)})]\mathbb{I}_{\{t\le \max\{D_i T_i\}\}} = \widehat{F}_{KM}(t)\mathbb{I}_{\{t\le \max\{D_i T_i\}\}}$$

   where $\max\{D_i T_i\}$ represents the largest uncensored duration.
3. If there are *n*o ties at $T_{(i)}$, then:

$$B(T_{(i)}) = D_i^{'}, \quad R(t_{(i)}) = n-i+1,$$

$$\widehat{S}_{KM}(t) = \prod_{\{T_{(i)}\le t\}} \left(1 - \frac{D_i}{n-i+1}\right)$$

   In many data sets, ties are observed, as a matter of fact. They call for two remarks: (i) even if F_η and F_ζ are continuous, $\Pr(\eta = \zeta) > 0$ is possible when η is not independent of ζ (see, for instance, Marshall and Olkin (1967)); (ii) the rounding problem: although theoretical models assume the time is continuous, actual measurements are discrete in nature. We have just seen that the Kaplan–Meier

estimator accommodates for ties. When the rounding problem is too severe because spells are actually observed through intervals, truncated survivor functions may be used for an explicit modelling.

4. If, at the largest observation, some censored *and* uncensored data are tied, the estimated distribution, $\hat{S}_{KM}(T_{(i)})$, is again defective and discontinuous at $T_{(n)}$, with:

$$\hat{S}_{KM}(T_{(n)}) > \hat{S}_{KM}(T_{(\infty-)}) > 0$$

17.4.3.2 Conditional Models: The Semi-parametric Proportional Hazard Model (The Cox Model)

Remember that in $\theta = (\alpha, \beta)$, α is a sufficient parameter for the baseline distribution, whereas β is introduced for describing the action of the exogenous variables. The semiparametric version of the proportional hazard model takes the form:

$$\lambda_T(t \mid z, \theta) = \alpha(t) \exp\left(z'\beta\right)$$

where $\alpha(t) = \lambda_0(t|z, \theta)$, which is the baseline hazard function, is now a functional parameter. Thus the parameter space takes the following form:

$$\begin{aligned}
\theta &= (\alpha, \beta) \in \Theta_\alpha \times \Theta_\beta \\
\Theta_\alpha &= \{\alpha : \mathbb{R}_+ \to \mathbb{R}_+ \mid \alpha \text{ is continuous and } \int_0^\infty \alpha(t)\mathrm{d}t = \infty\} \\
\Theta_\beta &\subset \mathbb{R}^k
\end{aligned}$$

The functional parameter α is often a nuisance parameter, whereas the Euclidean parameter β is the parameter of interest. It is therefore important to try to separate inferences on α and β. A natural idea is to construct a statistic $W = f(Y)$ such that the likelihood function $L_{Y|Z}(\alpha, \beta)$ factorizes as follows:

$$L_{Y|Z}(\alpha, \beta) = L_{W|Z}(\beta) \times L_{Y|W,Z}(\alpha, \beta)$$

In such a case, the inference on β would be made simpler by considering only the partial likelihood $L_{W|Z}(\beta)$ instead of $L_{Y|Z}(\alpha, \beta)$. A heuristic argument in favour of this simplification is that the information on β contained in $L_{Y|W,Z}(\alpha, \beta)$ is likely to be "eaten up" by the functional parameter α. This simplified estimator may now be build as follows. Similarly to the Kaplan–Meier estimator, let us reorder the sample according to the observed durations:

$$\begin{aligned}
T_i &\longrightarrow T_{(1)} < T_{(2)} < \ldots < T_{(n)} \\
D_i &\longrightarrow D_1', D_2', \ldots, D_n'
\end{aligned}$$

and let us also define:

$$R(t) = \sum_{1 \le i \le n} \mathbb{I}_{\{T_{(i)} \ge t\}}$$

$$\mathscr{R}(t) = \{k | T(k) \ge t\} = \{i | T_i \ge t\}$$

Thus $R(t)$ represents the number of individuals at risk at time t and $\mathscr{R}(t)$ represents the set of such individuals. Notation will be usefully simplified as follows:

$$R_{(i)} = R(T_{(i)}), \qquad \mathscr{R}_{(i)} = \mathscr{R}(T_{(i)})$$

Let us now represent the sample $(T_1, \ldots, T_n)$ by its order statistics $(T_{(1)} \ldots T_{(n)})$ and its rank statistics $(R_1, \ldots, R_n)$ where R_i is the rank of the i-th observation in the vector of order statistics. Giving the rank statistics, which plays the role of W in the previous expression, we may write the likelihood function of the rank statistics as follows:

$$L(\beta) = \prod_{1 \le i \le n} \left[\frac{e^{z_i' \beta}}{\sum_{k \in \mathscr{R}(i)} e^{z_k' \beta}} \right]^{D_i} = \prod_{1 \le i \le D_+} \left[\frac{e^{z_i' \beta}}{\sum_{k \in \mathscr{R}(i)} e^{z_k' \beta}} \right]^{D_i}$$

where $D_+ = \sum_i D_i$. The (partial) likelihood estimator of β is then defined as

$$\hat{\beta} = \arg\max_{\beta} L(\beta)$$

This estimator is consistent and its asymptotic properties have been studied e.g. by Tsiatis (1981) and by Andersen, Borgan, Gill and Keiding (1993).

17.5 Counting Processes and Point Processes

Point processes provide the framework for modelling trajectories with more than one transition and more than two states (such trajectories are sometimes called duration models with multiple spells and multiple states). Formally a point process is a continuous time process with a finite state space and right continuous with left limit (*cadlag*) trajectories. A point process is conveniently represented by means of a multivariate counting process that counts, as time increases, the number of possible transitions. Consequently, we will first present counting processes.

17.5.1 Definitions

Let us consider a (finite or infinite) sequence $(T_p)_{p \ge 1}$ of increasing random durations $(0 < T_1 < T_2 < \ldots)$. This sequence characterizes a univariate counting process:

$$N_t = \sum_{p \ge 1} \mathbb{I}(T_p \le t)$$

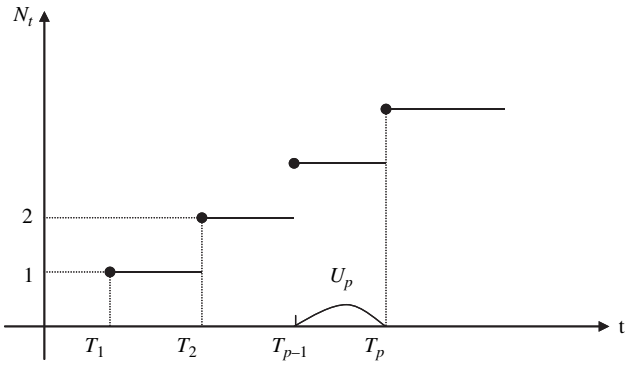

Fig. 17.2 A realization of a univariate counting process

The trajectories of N_t are right continuous, and such that $N_0 = 0$ and N_t only increases by jumps of size 1. A typical realization is shown in Fig. 17.2.

A duration model defines a process with a single jump ($N_t = \mathbb{I}(T \geq t)$). From the definition of N_t, we can deduce easily the definition of the date T_p of the j-th jump of the process:

$$T_p = \inf\{t | N_t = p\}, \quad p \geq 1$$

The distribution of N_t may be characterized by the distribution of the sequence $(T_p)_{p\geq 1}$. Equivalently, that sequence may be replaced by the sequence of positive numbers:

$$U_p = T_p - T_{p-1} \quad (T_0 = 0)$$

The random variable is now the duration between the $(p-1)$-th and the p-th jumps. If the random variables $(U_p)_{p\geq 1}$ are i.i.d., the process is called a *renewal process*. The information denoted by $\mathscr{F}_t^N$ and carried by N_t, observed from 0 to t (included), is equivalent to the knowledge of $T_1, \ldots, T_p$ $(T_p \leq t < T_{p+1})$ and the event $T_{p+1} > t$. Equivalently this information may be described by the random variables $U_1, \ldots, U_p$ and by the event $U_{p+1} > t - \sum_{q=1}^{p} U_q$.

A *multivariate counting process* is a vector $N_t = (N_t^1, \ldots, N_t^J)$ of counting processes. This vectorial process is characterized by J sequences $(T_p^j)_{p\geq 1}$ $(j = 1, \ldots, J)$ of increasing durations and by:

$$N_t^j = \sum_{p\geq 1} \mathbb{I}(T_p^j \leq t)$$

The information content of the observation of this multivariate counting process up to time t is described by the family of random variables T_p^j such that $T_{p_j}^j \leq t$, and by the J events $T_{p_j+1}^j > t$.

A multivariate counting process may also be represented by a unique sequence $(T_r)_{r\geq 1}$ of the jump times of any element of the vector N_t, and by e_r $(r \geq 1)$ which is a discrete-time process valued in $(1, \ldots, J)$. In this sequence $(T_r, e_r)_{r\geq 1}$, e_r ; indicates the component j that jumps at date T_r. Note that the sequence (T_r) has the property:

$$\bar{N}_t = \sum_{j=1}^{J} N_t^j = \sum_{r\geq 1} \mathbb{I}(T_r \leq t)$$

The distribution of N_t may then be described by the sequence of conditional distributions:

$$(T_r, e_r) \mid (T_s, e_s)_{s=1,\ldots,r-1}$$

Consider for example a bivariate duration (T^1, T^2), where $\Pr(T^1 = T^2) = 0$. This pair defines two single jump counting processes:

$$N_t^1 = \mathbb{I}(T^1 \leq t) \quad \text{and} \quad N_t^2 = \mathbb{I}(T^2 \leq t)$$

Then the $(T_r)_r$ sequence becomes:

$$T_1 = \min(T^1, T^2), \quad T_2 = \max(T^1, T^2)$$

and

$$\begin{aligned} e_1 &= \mathbb{I}(T^1 < T^2) + 2\mathbb{I}(T^2 \leq T^1) \\ e_2 &= 3 - e_1 \end{aligned}$$

A point process is a continuous-time process valued in a finite (or more generally discrete) state space $\{1,\ldots,K\}$. Such a process X_t may represent, for example, the labor market situation of an individual at time t. In such a case, the set $\{1,\ldots,K\}$ describes the different possible labor market states (full-time employed, part-time employed, unemployed, retired,...) and X_t is characterized by the dates of the transitions between two different states. Indeed, a point process defines a multivariate counting process. Consequently, we denote by $j = (k,k')$ the pair of states such that a transition from k to k' is possible and $\{1,\ldots,J\}$ is the set of all these ordered pairs. Then $(T_p^j)_{p\geq 1}$ is the sequence of jump times from k to k' if $j = (k,k')$ and

$$N_t^j = \sum_{p\geq 1} \mathbb{I}(T_p^j \leq t)$$

This multivariate counting process satisfies the following constraint by construction: after a jump of the component N_t^j, $j = (k,k')$, the next process which may jump is necessarily an element of the subfamily $(N_t^\ell)_\ell$ where $\ell = (k', k")$ and $k" \neq k'$.

17.5.2 Stochastic Intensity, Compensator and Likelihood of a Counting Process

The stochastic intensity of a univariate counting process is defined as follows:

$$h_N(t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \Pr(N_{t+\Delta t} - N_t = 1 \mid \mathscr{F}_{t-}^N)$$

If for instance $N_t = \mathbb{I}(T \leq t)$, this definition implies, where T is a continuous variable, that $h(t) = \lambda(t)$, which is the hazard function of T if $T > t$ and $h(t) = 0$ after the jump T. Equivalently:

$$h_N(t) = \lambda_T(t)(1 - N_{t-}),$$

where $N_{t-} = \mathbb{I}(T < t)$.

If N_t is a general univariate counting process $N_t = \sum_{p \geq 1} \mathbb{I}(T_p \geq t)$, the stochastic intensity is obtained by the following rule:

- If $t > \max_p(T_p)$ then $h(t) = 0$
- If t verifies $T_{p-1} < t \leq T_p$ (where $p = N_t + 1$) then

$$h_N(t) = \lambda_p(t \mid T_1, \ldots, T_{p-1})$$

where λ_p is the hazard function of the duration T_p conditional on $T_1, \ldots, T_{p-1}$.

If the model is specified in terms of $U_p = T_p - T_{p-1}$, we have

$$h_N(t) = \lambda_p^U(t - T_{p-1} \mid U_1, \ldots, U_{p-1})$$

where λ_p^U is the hazard function of U_p given $U_1, \ldots U_{p-1}$. This definition is easily extended to multivariate counting processes. The stochastic intensity is then multivariate and for each $j \in \{1, \ldots, J\}$:

$$h_N^j(t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \Pr(N_{t+\Delta t}^j - N_t^j = 1 \mid \mathscr{F}_{t-}^N)$$

where $\mathscr{F}_{t-}^N$ represents the information carried by all the coordinates of the process observed before t.

If $N_t^j = \sum_{p \geq 1} \mathbb{I}(T_p^j \leq t)$, $h_N^j(t)$ is null if $t > \max_p(T_p^j)$. For each coordinate ℓ, we can choose $p_\ell = N_t^\ell + 1$ such that

$$T_{p_\ell - 1}^\ell \leq t < T_{p_\ell}^\ell$$

(where $T_{p_\ell}^\ell = +\infty$ if N_t^ℓ never jumps after $T_{p_\ell - 1}^\ell$) and $\lambda_T^i(t)$ is equal to the hazard function of $T_{p_j}^j$ at the point t, given all the T_q^ℓ, $\ell \neq j$ and $q < p_\ell$, and the family of events $T_{p_\ell}^\ell \geq t$. Let us take as an example the bivariate counting process $N_t^1 = \mathbb{I}(T^1 \leq t)$, $N_t^2 = \mathbb{I}(T^2 \leq t)$. The stochastic intensity $h_N^1(t)$ is equal to the hazard function of T^1 conditional on $T^2 = t_2$ if $T^2 < t$ or conditional on $T^2 \geq t$ if $T^2 \geq t$. The compensator of univariate counting process N_t with stochastic intensity $h_N(t)$ is defined by

$$H_N(t) = \int_0^t h(s) d_s$$

For a duration model $N_t = \mathbb{I}(T \leq t)$, $H_N(t)$ is equal to the integrated hazard $\Lambda(t)$ if $T > t$ and equal to $\Lambda(T)$ if $T \leq t$.

For a multivariate counting process N_t^j, we define a vector of compensators by:

$$H_N^j(t) = \int_0^t h_N^j(s)\mathrm{d}s\ .$$

From now on, we simplify the notation $H_N^j(t)$ into H_t^j similarly to N_t instead of $N(t)$. The compensators are positive and non-decreasing predictable processes satisfying $H_0 = 0$. The main property of the compensator is that the difference:

$$M_t = N_t - H_t$$

is a zero mean $\mathscr{F}_t^N$- martingale (i.e. $E(M_t \mid \mathscr{F}_s^N) = M_s$). The decomposition $N_t = H_t + M_t$ is called the Doob–Meyer decomposition of the process N_t. The same decomposition may be constructed for a multivariate counting process. In that case, M_t^j is a martingale with respect to the information sets generated by the whole process $(N_t^1, \ldots, N_t^J)$.

The stochastic intensity and the compensator both determine an elegant expression of the likelihood of a counting process. Consider first a univariate process $N_t = \sum_{p\geq 1} \mathbb{I}(T_p \leq t)$. If the process is observed between 0 and t such that $T_{p-1} < t < T_p$, the likelihood of this observation is:

$$\ell(t) = \left\{ \prod_{q=1}^{p-1} f_q(T_q \mid T_1, \ldots, T_{q-1}) \right\} \times S_p(t \mid T_1, \ldots, T_{p-1})$$

where f_q and S_q are respectively the density and the survivor functions of T_q given $T_1, \ldots, T_{q-1}$. One can easily check that:

$$\ell(t) = \prod_{T_q \leq t} h(T_q) e^{-H_t}$$

or

$$\ln \ell(t) = \int_0^t \ln h(s)\mathrm{d}N_s - H_t$$

In this expression, we use the stochastic integral notation:

$$\int_0^t g(s)\mathrm{d}N_s = \sum_{T_p \leq t} g(T_p)$$

The stochastic intensity notation can be generalized to multivariate processes for which the likelihood corresponding to the observation of all the coordinates of the process up to time t is equal to:

$$\ln l(t) = \sum_{j=1}^{J} \left\{ \int_0^t \ln h^j(s)\mathrm{d}N_s^j - H_t^j \right\}$$

This way of writing the likelihood function is the basis for Cox's estimation and martingale estimations, to be presented in the last section of this chapter.

17.6 Poisson, Markov and Semi-Markov Processes

In this section, we give first the example of a well-known single counting process, namely the Poisson process. Then we examine point processes displaying Markovian or semi-Markovian properties.

17.6.1 Poisson Processes

We consider the familiar Poisson process as an example of a univariate counting process. Let M be a positive measure on $\mathbb{R}^+$ with density m with respect to the Lebesgue measure, i.e., $M([a,b]) = \int_a^b m(x)\, d(x)$.

A stochastic process N_t is a *Poisson process* associated with the measure M if its distribution satisfies the following requirements:

(i) $N_0 = 0$,
(ii) N_t is a process with independent increments: $\forall t_1, \ldots, t_n$, the random variables $(N_{t_i} - N_{t_{i-1}})_{i=1,\ldots,n}$ are independent random variables,
(iii) the distribution of $(N_t - N_s)$ is a Poisson distribution for any $s < t$, which means that:

$$\Pr(N_t - N_s = k) = \frac{M([s,t])^k}{k!} e^{-M([s,t])}$$

These three properties imply that a Poisson process is a counting process with unit jumps. If $m(x)$ is equal to some positive constant λ, then the process is said to be homogeneous and we may verify that sojourn times $U_p = T_n - T_{n-1}$ are i.i.d. random variables with an exponential distribution with parameter $\lambda > 0$. The *homogeneous Poisson process* is then the renewal process characterized by the exponential distribution.

The compensator and the intensity of a Poisson process, with respect to its canonical filtration, are equal to $H = M([0,t])$ and to $m(t)$, respectively. This result follows from the equalities:

$$\begin{aligned} h(t) &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \Pr(N_{t+\Delta t} - N_t = 1 \mid \mathscr{F}_{t^-}^N) \\ &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \Pr(N_{t+\Delta t} - N_t = 1 \mid N_t) \\ &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} [M(t, t+\Delta t]) e^{-M([t,t+\Delta t])}] \\ &= m(t) \end{aligned}$$

In particular, if the process is homogeneous, $h(t)$ is constant. The likelihood $\ell(t)$ relative to the observation of the process N_t between 0 and t is derived from the intensity and the compensator, i.e.

$$\begin{aligned}\ln \ell(t) &= \int_0^t [\ln m(s)]\, \mathrm{d}N_s - M([0,t]) \\ &= \sum_{\tau_n \le t} \ln m(\tau_n) - M([0,t])\,.\end{aligned}$$

If N_t is an homogeneous Poisson process with parameter λ, its likelihood satisfies:

$$\ln \ell(t) = N_t \ln \lambda - \lambda t\ .$$

17.6.2 Markov Processes

17.6.2.1 Definitions

We consider a point process $X = (X_t)_{t\in\mathbb{R}^+}$ valued in the finite state space $E = \{1,\ldots,K\}$. The distribution of X_t is totally defined by a projective system:

$$\Pr(X_{t_1} = j_1, \ldots, X_{t_p} = j_p)$$

for any finite subset $(t_1,\ldots,t_p)$ of $\mathbb{R}^+$ satisfying $t_1 < t_2 < \cdots < t_p$. From these probabilities, one can compute:

$$\Pr(X_{t_p} = j_p \mid X_{t_1} = j_1, \ldots, X_{t_{p-1}} = j_{p-1})$$

and the process X_t is a *Markov process* if:

$$\Pr(X_{t_p} = j_p \mid X_{t_1} = j_1, \ldots, X_{t_{p-1}} = j_{p-1}) = \Pr(X_{t_p} = j_p \mid X_{t_{p-1}} = j_{p-1})$$

It follows that a Markov process is characterized by the distribution of the initial condition, i.e. by the distribution of X_0, and by the *transition probabilities*:

$$p_{jk}(s, s+t) = \Pr(X_{s+t} = k \mid X_s = j)$$

defined for any s and $t \in \mathbb{R}^+$, and for any j and $k \in E$. The Markov process is said to be *time-homogeneous* if:

$$p_{jk}(s, s+t) = p_{jk}(0,t), \quad \forall (s,t) \in \mathbb{R}^+ \times \mathbb{R}^+, \forall (j,k) \in E^2\ ,$$

i.e. if the transition probability does not depend on the origin of the time set, but only on the difference between the two dates s and $(s+t)$. For a time-homogeneous Markov process, we denote the transition probability $p_{jk}(0,t)$ by $p_{jk}(t)$ and the matrix with elements $p_{jk}(t)$ by $P(t)$. So, $P(t)$ is a $K \times K$ matrix of non-negative numbers such that the sum of each row is equal to one, i.e.

$$\sum_{k=1}^{K} p_{jk}(t) = 1$$

Moreover, decomposing the trajectory on $[0,t]$ into two sub-trajectories on $[0,s]$ and $[s,t]$, we obtain the following properties of the matrices $P(t)$:

$$p_{jk}(t) = \sum_{\ell=1}^{K} p_{j\ell}(s)\, p_{\ell k}(t-s)\,,\ \forall\ 0 \le s \le t,\ \forall\ (j,k) \in E \times E$$

or equivalently:

$$P(t) = P(s)\, P(t-s)\,, 0 \le s \le t\ .$$

We will now restrict our attention to processes satisfying some regularity conditions.

Definition 17.8. A time-homogeneous Markov process X_t is said to be *standard* if:

(i) $\forall j \in E,\ \lim_{t \downarrow 0} p_{jj}(t) = 1$, and then, $\quad \forall k \neq j,\ \lim_{t \downarrow 0} p_{jk}(t) = 0,$

(ii) $\exists q_{jk} \in \mathbb{R}^+, \forall (j,k) \in (E \times E)$, with $k \neq j$,

$q_{jk} = \lim_{t \downarrow 0} \frac{1}{t} p_{jk}(t) = \frac{d}{\mathrm{d}t} p_{jk}(t)\ |_{t=0},$

$q_{jj} = -\lim_{t \downarrow 0} \frac{1}{t}(1 - p_{jj}(t)) = -\sum_{k \neq j} q_{jk}$ ■

As a direct consequence, quantities q_{jk} satisfy the following properties:

$$\begin{array}{lll} \text{(i)} & \sum_{k=1}^{K} q_{jk} = 0, & j \in E, \\ \text{(ii)} & q_{jk} \ge 0, \quad k \neq j, \quad \text{and} \quad q_{jj} \le 0, & j \in E\ . \end{array}$$

If $j \neq k$, q_{jk} is called the *intensity of transition* from state j to state k. The matrix Q is called the *intensity matrix* or the *generator* of the process X_t. Writing p_{jj} as $p_{jj}(t) = 1 - \sum_{k \neq j} p_{kj}(t)$, the previous definition implies that $Q = \dfrac{d}{\mathrm{d}t} P(t)|_{t=0}$

Theorem 17.2. *The transition matrix $P(t)$ of the time-homogeneous standard Markov process X_t satisfies the forward matrix equation*

$$\frac{d}{\mathrm{d}t} P(t) = P(t) \cdot Q$$

and the backward matrix equation

$$\frac{d}{\mathrm{d}t} P(t) = Q \cdot P(t)\ .$$

Proof. See Doob (1953), pp. 240–241, or Bhattacharya and Waymire (1990), pp. 263–267.

These two equations are known as the *Kolmogorov forward and backward differential equations*, respectively. In general, these equations do not have a unique solution; however, X_t is said to be regular if the solution, subject to the border condition $P(0) = I$, is unique and has the exponential form given in the following theorem (where I is the identity matrix).

Theorem 17.3. *If the time-homogeneous Markov process X_t with generator Q is regular, then the matrix*

$$P(t) = \exp(Qt) = \sum_{n=0}^{\infty} t^n Q^n / n! \tag{17.1}$$

exists for any t, and is the unique solution to the Kolmogorov differential equations subject to the border condition $P(0) = I$.

Proof. See Doob (1953), pp. 240–241, or Bhattacharya and Waymire (1990), pp. 267–275.

17.6.2.2 Distributions Related to a Time-Homogeneous Standard Markov Process

Since the state space E is finite, the Markov process X_t moves by jumping from one state to another. Let $0 = T_0 < T_1 < T_2 < \cdots$, be the times of these transitions. As the sample paths of the process X_t are right-continuous step functions, we can define $Y_n = X_{T_n}$ as the state entered at T_n. Moreover, we set:

$$U_n = T_n - T_{n-1}\,,\, n \in \mathbb{N}, \text{ and } U_0 = 0$$

The random variable U_n represents the sojourn duration of the process in state $Y_{n-1} = X_{T_{n-1}}$ entered at time T_{n-1}. A Markov point process X_t can be represented by a multivariate counting process characterized by the sequence $(T_n, e_n)_{n\geq 0}$. In this representation, e_n is the transition at time T_n, i.e.:

$$e_n = (Y_{n-1}, Y_n) \text{ with } Y_{n-1} \neq Y_n\,.$$

Thus e_n takes its value in a finite set with $K(K-1)$ elements. Yet, the representation of X_t as a point process is easier to formalize. So, we are interested in the distribution of the sequences $(T_n, Y_n)_{n\geq 0}$ or $(U_n, Y_n)_{n\geq 0}$, rather than of the sequence $(T_n, E_n)_{n\geq 0}$.

For that purpose, we firstly set $\lambda_j = -q_{jj}$ for any $j \in E$, and we define quantities ρ_{jk} as follows:

- If $\lambda_j \neq 0$, $j \in E$, $\quad \rho_{jj} = 0$ and $\rho_{jk} = q_{jk}/\lambda_j\,,\, k \neq j$
- If $\lambda_j = 0\,,\, j \in E$ $\quad \rho_{jj} = 1$ and $\rho_{jk} = 0\,,\, k \neq j$

Theorem 17.4. *If X_t is a time-homogeneous standard Markov process, then*

(i) $(U_n, Y_n)_{n\geq 0}$ is a Markov sequence and (U_n, Y_n) is independent of U_{n-1} given Y_{n-1}. Moreover U_n and Y_n are conditionally independent given Y_{n-1}.
(ii) U_n given $Y_{n-1} = j$ has an exponential distribution with parameter λ_j if $\lambda_j \neq 0$. If $\lambda_j = 0$, the state j is absorbing and $U_n = \infty$ with probability 1.
(iii) $Y = (Y_n)_{n\geq 0}$ is a Markov chain with transition matrix:

$$\Pr(Y_n = k \mid Y_{n-1} = j) = \rho_{jk},\ (j,k) \in E \times E$$

Proof. See Bhattacharya and Waymire (1990), pp. 275–279.

Theorem 17.5. *If X_t is irreducible $(\forall j, \forall k, \exists m$ such that $p_{jk}(m) > 0)$ and recurrent $(\Pr(\inf\{m \mid Y_{n+m} = j\} < \infty \mid Y_n = j) = 1)$ then:*

(i) the limits of transition probabilities $p_{jk}(t)$ exist and are independent of the initial state, i.e.

$$\lim_{t \uparrow \infty} p_{jk}(t) = \Pi_k$$

(ii) either $\Pi = (\Pi_1, \ldots, \Pi_K) = (0, \ldots, 0)$, in which case all states are said to be null recurrent, or $\sum_{k=1}^{K} \Pi_k = 1$, in which case all states are said to be non-null recurrent (or positive recurrent if $\Pi_k > 0, \forall k \in E$).

Proof. See Cox and Miller (1966), pp. 106–117.

The *limiting distribution* Π is also invariant or stationary, because:

$$\Pi = \Pi \cdot P(t)\,, \forall t \in \mathbb{R}^+$$

In the case of an irreducible, recurrent non-null Markov process with generator Q, calculation of the vector Π is made easier by noting that Π is the unique invariant distribution probability satisfying the linear equation:

$$\Pi \cdot Q = 0$$

Moreover, if the embedded Markov chain Y is also irreducible and recurrent non-null, Y has a limit distribution v satisfying:

$$v = v \cdot R$$

where R is the transition matrix of the embedded Markov chain. The relationship between the two limit distributions Π and v is:

$$\Pi_j = \left[\frac{v_j}{\lambda_j}\right] \left[\sum_{k=1}^{K} \frac{v_k}{\lambda_k}\right], \qquad j \in E$$

or equivalently:

$$v_j = \frac{\Pi_j \lambda_j}{\left[\sum_{k=1}^{K} \Pi_k \lambda_k\right]}, \qquad j \in E$$

Generally, v and Π are different. The last equation has a very interesting interpretation: since v_j is the long-run frequency of visits of the chain Y to state j, and since $(1/\lambda_j)$ is the mean duration of a sojourn of the process X_t in state j, then Π_j, which is the long-run proportion of occupation of state j for the process x_t, is calculated as the long-run global duration of sojourn in state j for the process X_t (calculated as

the product of v_j and $1/\lambda_j$), divided by the sum of the long-run global durations of sojourn in the different states.

17.6.2.3 Statistical Inference for Time-Homogeneous Markov Models

Now we shall discuss the problem of estimating the generator Q of a time-homogeneous Markov process X_t from the observation of N independent sample paths over a fixed time interval $[0,T]$. Firstly, we consider the nonparametric case of N i.i.d. realizations of X_t over $[0,T]$. Here the likelihood function $L_Q^{(N)}$ is given by

$$L_Q^{(N)} = \prod_{i=1}^{N} \left\{ \Pr(x_0^{(i)} = Y_0^{(i)}) \times e^{-(T-\tau_{n_i}(i))\lambda_{y_{n_i}}(i)} \times \prod_{j=0}^{n_i-1} q_{Y_j^{(i)},Y_{j+1}^{(i)}} \, e^{-u_{j+1}^{(i)}\lambda_{Y_j^{(i)}}} \right\}$$

where n_i is the number of transitions observed for the i–th sample path over $[0,T]$, $0 < \tau_1^{(i)} < \tau_2^{(i)} < \cdots < \tau_{n_i}^{(i)}$ being the ordered sequence of transition times for this sample, and $\{(u_0^{(i)}, Y_0^{(i)}), (u_1^{(i)}, Y_1^{(i)}), \ldots, (u_{n_i}^{(i)}, Y_{n_i}^{(i)})\}$ being the sequence of successive sojourn durations and visited states for the i–th sample path, with the conventions: $u_0^{(i)} = 0$ and $u_{n_i+1}^{(i)} = T - \tau_{n_i}^{(i)}$.

If we denote $\mathcal{N}_{T,N}(j,k)$ the total number of transitions from state j to state k observed over the N realizations and $D_{T,N}(j)$ the total length of time that state j is occupied during these N sample paths, then it is easy to show that the maximum likelihood estimator for $q_{j,k}$ is given by:

$$\widehat{q}_{j,k}(T,N) = \frac{\mathcal{N}_{T,N}(j,k)}{D_{T,N}(j)}$$

if $j \neq k$ and $D_{T,N}(j) \neq 0$. If $D_{T,N}(j) = 0$, the MLE of $q_{j,k}$ does not exist and we adopt the convention that:

$$\widehat{q}_{j,k}(T,N) = 0 \text{ if } j \neq k \text{ and } D_{T,N}(j) = 0\,.$$

Asymptotic properties of the MLE estimates $\widehat{q}_{j,k}(T,N)$ when $T \to \infty$ and N is fixed (typically, $N = 1$), or when $N \to \infty$ and T is fixed, are given by the following theorems (see Basawa and Prakasa Rao (1980), pp. 195–197).

Theorem 17.6. *If there is a positive probability that the j-th state will be occupied at some date $t \geq 0$, then*

$$\operatorname*{plim}_{T\uparrow\infty} \widehat{q}_{j,k}(T,N) = q_{j,k} \qquad a.s.$$

and if each state has a positive probability of being occupied, then the random variables

$$\left\{N^{1/2}(\widehat{q}_{j,k}(T,N)-q_{j,k})\right\}_{k\neq j}$$

are asymptotically normal and independent with zero mean and variance

$$\frac{q_{j,k}}{\int_0^T \Pr[X_t=j]\,\mathrm{d}t}$$

Moreover, $\widehat{q}_{j,k}(T,N)$ is asymptotically efficient when N tends to infinity. ■

Theorem 17.7. *If the time-homogeneous standard Markov process X_t is regular and recurrent positive, then*

$$\operatorname*{plim}_{T\uparrow\infty} \widehat{q}_{j,k}(T,1)=q_{j,k} \qquad a.s.$$

and the random variables

$$\{T^{1/2}(\widehat{q}_{j,k}(T,1)-q_{j,k}\}_{j,k=1,\ldots,K,\ j\neq k}$$

are asymptotically normal and independent with zero mean and variance $q_{j,k}\rho/\overline{Q}(j,j)$ where ρ is the product of the non-zero eigenvalues of Q and $\overline{Q}(j,j)$ is the (j,j)-th cofactor of Q. ■

In the last case (i.e. when $N=1, T\uparrow\infty$), it is shown that

$$\int_0^T \Pr[x_t=j]\mathrm{d}t=\overline{Q}(j,j)T\rho^{-1}+o(T) \quad \text{for } T\uparrow\infty$$

Since

$$\frac{1}{T}E\left\{\frac{\partial \mathrm{Log}L_Q^{(1)}}{\partial q_{j,k}}\right\}^2=\frac{\int_0^T \Pr[x_t=j]\,\mathrm{d}t}{Tq_{j,k}}$$

it follows from the previous theorem that

$$T^{1/2}\left\{\widehat{q}_{j,k}(T,1)-q_{j,k}\right\}\xrightarrow{d} N\left[0,q_{j,k}\rho/\overline{Q}(j,j)\right]$$

and so $\widehat{q}_{j,k}(T,1)$ is asymptotically efficient for $T\uparrow\infty$.

Now let us suppose that transition intensities are functions of a set $\theta=(\theta_1,\ldots,\theta_p)$ of unknown parameters, i.e. they have the form $q_{j,k}(\theta)$. The problem is then to obtain a MLE of θ from N independent observations of the process $x_t(\theta)$ over the period $[0,T]$. In this case, the likelihood function is:

$$
\begin{aligned}
L_Q(N,\theta) &= \prod_{i=1}^{N} \left\{ \Pr\left(x_0(i,\theta) = Y_0^{(i)}\right) \times e^{-(T-\tau_{n_i}^{(i)})\lambda_{Y_{n_i}^{(i)}}^{(\theta)}} \right. \\
&\quad \left. \times \prod_{j=0}^{n_i-1} q_{Y_j^{(i)},Y_{j+1}^{(i)}}^{(\theta)} \, e^{-u_{j+1}^{(i)}\lambda_{Y_j^{(i)}}^{(\theta)}} \right\} \\
&= \left\{ \prod_{i=1}^{N} \Pr(x_0(i,\theta) = Y_0^{(i)}) \right\} \times \left\{ \prod_{j,k=1;j\neq k}^{K} (q_{j,k}^{(\theta)})^{\mathscr{N}_{T,N}(j,k)} \right\} \\
&\quad \times \left\{ \prod_{j=1}^{K} e^{-\lambda_j^{(\theta)} D_{T,N}(j)} \right\}
\end{aligned}
$$

where:

- $D_{T,N}(j) = \sum_{i=1}^{N} \sum_{\ell=0}^{n_i} u_{\ell+1}^{(i)} \mathbb{I}(Y_\ell^{(i)} = j)$ is the total sojourn duration in state j, which is observed over the N sample paths (with the convention $u_{n_i+1}^{(i)} = T - \tau_{n_i}^{(i)}$),
- $\mathscr{N}_{T,N}(j,k) = \sum_{i=1}^{N} \sum_{\ell=0}^{n_i-1} \mathbb{I}\left(Y_\ell^{(i)} = j, Y_{\ell+1}^{(i)} = k\right)$ is the total number of transitions from j to k, observed over the N sample paths.

With the assumption that the initial state $Y_0^{(i)}$ does not depend on θ, the ML equations for estimating θ are:

$$
\frac{\partial \operatorname{Log} L_Q(N,\theta)}{\partial \theta_m} = \prod_{\substack{j,k=1 \\ j\neq k}}^{K} \left[\frac{\mathscr{N}_{T,N}(j,k) - D_{T,N}(j) q_{jk}^{(\theta)}}{q_{jk}^{(\theta)}} \right] \frac{\mathrm{d}q_{jk}^{(\theta)}}{\mathrm{d}\theta_m} = 0, \; m = 1, \ldots, \; p
$$

In the case where $N = 1$, Billingsley (1961), p. 46, has shown that these equations yield a consistent solution $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p)$ such that

$$
\mathscr{N}_T^{1/2}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, i(\theta)^{-1}) \text{ as } T \uparrow \infty
$$

where $\mathscr{N}_T$ is the total number of transitions during the interval $[0,T]$ and

$$
\begin{aligned}
i(\theta) &= -\frac{1}{\mathscr{N}_T} \left[E\left(\frac{\partial^2 \operatorname{Log} L_Q(\theta)}{\partial \theta_m \, \partial \theta_m'} \right) \right]_{m=1,\ldots,p} \\
&= \left[\prod_{\substack{j,k=1 \\ j\neq k}}^{K} \frac{v_j(\theta)}{\lambda_j(\theta) \, q_{jk}^{(\theta)}} \left(\frac{\mathrm{d}q_{jk}^{(\theta)}}{\mathrm{d}\theta_m} \right) \left(\frac{\mathrm{d}q_{jk}^{(\theta)}}{\mathrm{d}\theta_{m'}} \right) \right]_{m, \, m'=1,\cdots,p}
\end{aligned}
$$

$v(\theta) = [v_j(\theta)]_{j=1,\ldots,K}$ being the limit distribution of the embedded Markov chain associated to the process X_t.

17.6.3 Semi-Markov Processes

Semi-Markov processes can be viewed as direct extensions of Markov processes. Using notations and concepts introduced for the characterization of a Markov process, we will say that a stochastic process $\{X_t\}_{t\geq 0}$ taking its value in the discrete state space $E = \{1,\ldots,K\}$ is semi-markovian if the sequence $\{Y_n\}_{n\geq 0}$ of states visited remains a Markov chain, but time u_n spent in the state Y_{n-1} need not be exponentially distributed and may depend on the next state entered, namely Y_n.

Definition 17.9. If $(Y_n)_{n\in\mathbb{N}}$ and $(u_n)_{n\in\mathbb{N}}$ denote respectively the sequences of visited states and sojourn durations of a continuous-time process $\{X_t\}_{t\geq 0}$ with a discrete state space $E = \{1,\ldots,K\}$, then $\{X_t\}_{t\geq 0}$ is a *semi-Markov process* if:

$$\begin{aligned}\Pr\ &\left\{Y_\ell = j, u_\ell \leq t \mid (Y_n)_0^{\ell-1}, (u_n)_0^{\ell-1}\right\}\\ &= \Pr\{Y_\ell = j, u_\ell \leq t \mid Y_{\ell-1}\} \qquad \ell \in \mathbb{N}, j \in E, t \in \mathbb{R}^+\end{aligned}$$

with the convention $u_0 = 0$. Moreover, a semi-Markov process $\{X_t\}_{t\geq 0}$ is said to be *time-homogeneous* if transition probabilities

$$\Pr\{Y_\ell = j, u_\ell \leq t \mid Y_{\ell-1} = i\} = \mathscr{P}(i,j,t)\,,\ (i,j) \in E \times E$$

do not depend on ℓ. The function $\mathscr{P}$ is called the *kernel* of the semi-Markov process $\{x_t\}_{t\geq 0}$. Then the sequence $Y = (Y_n)_{n\in\mathbb{N}}$ is a Markov chain with transition matrix:

$$R(i,j) = \mathscr{P}(i,j,\infty) = \lim_{t\uparrow\infty} \mathscr{P}(i,j,t)\,,\ (i,j) \in E \times E$$

and $u_1, u_2, \ldots$ are conditionally independent given Y. ■

If the kernel $\mathscr{P}$ is defined as

$$\mathscr{P}(i,j,t) = \rho_{ij}(1 - e^{-\lambda_i t})\ ,\ (i,j) \in E \times E$$

where $\lambda_i \in]0,\infty[$, $\rho_{ii} = 0$ and $\sum_{j\in E} \rho_{ij} = 1$ $,\forall i \in E$, then $\{X_t\}_{t\geq 0}$ is a time-homogeneous Markov process with generator $Q(i,j) = q_{ij} = \lambda_i \rho_{ij},\ j \neq i$. On the other hand, if $E = \{i\}$ is a singleton, then $(u_n)_{n\in\mathbb{N}}$ is a time-homogeneous *renewal process* with an inter–arrival time distribution of the form $F(t) = \mathscr{P}(i,i,t)$.

The law of a semi-Markov process $\{X_t\}_{t\geq 0}$ is jointly characterized by the transition probability $R(i,j)$ of the embedded Markov chain $(Y_n)_{n\in\mathbb{N}}$ and the conditional sojourn distributions:

$$G(i,j,t) = \Pr\{u_\ell \leq t \mid Y_{\ell-1} = i, Y_\ell = j\}, \ell \in \mathbb{N}, (i,j) \in E \times E$$

The kernel function of this semi-Markov process is then defined as:

$$\mathscr{P}(i,j,t) = R(i,j)\ G(i,j,t)$$

from which are deduced unconditional sojourn distributions:

$$F(i,t) = \Pr\{u_\ell \le t \mid Y_{\ell-1} = i\} = \sum_{j \in E} \mathscr{P}(i,j,t)$$

Let us recall that if the Markov chain $Y = (Y_n)_{n\in\mathbb{N}}$ is irreducible and recurrent non-null, there exists a limiting probability distribution v on E of the form:

$$v_j = \sum_{i \in E} v_i R(i,j), \quad j \in E$$

or in matrix notation:

$$v = vR$$

Moreover, if $\overline{u}_1(i) = E[u_1 \mid Y_0 = i] < \infty$, $\forall i \in E$, then the limit distribution of the semi-Markov process $\{X_t\}_{t\ge 0}$ is given by:

$$\Pi_j = \lim_{t\uparrow\infty} \Pr\{x_t = j\} = \frac{v_j \overline{u}_1(j)}{\sum_{i\in E} v_i \overline{u}_1(i)}$$

Notice that this relation between Π and v is more general than the one for Markov processes, for which $\overline{u}_1(i) = \lambda_i^{-1}$.

The main statistical problem is to estimate the semi-Markov kernel $\mathscr{P}$. Here we concentrate on a fully nonparametric estimation procedure for a semi-Markov process $\{X_t\}_{t\ge 0}$, where the distribution of a sojourn in state i does not depend on the next state to be entered, i.e.:

$$G(i,j,t) = F(i,t), \quad \forall (i,j) \in E \times E, \forall t \in \mathbb{R}^+$$

Then $R(i,j)$ and $F(i,t)$ can be estimated from N i.i.d. realizations of $\{X_t\}_{t\ge 0}$ over a fixed time interval $[0,T]$. In that case, let us denote $\mathscr{N}_{T,N}(i,j)$ and $\mathscr{N}^*_{T,N}(i) = \sum_{j\in E} \mathscr{N}_{T,N}(i,j)$ the number of transitions from i to j in $[0,T]$ and the number of sojourns in state i completed before time T, respectively. Then nonparametric maximum-likelihood estimators of the unconditional sojourn distributions and of the transition matrix of the embedded Markov chain are respectively given by:

$$\widehat{F}(i,t) = \mathscr{N}^*_{T,N}(i)^{-1} \sum_{\ell=1}^{\mathscr{N}^*_{T,N}(i)} \mathbb{I}(Y_{\ell-1} = i, u_\ell \le t)$$

and

$$\widehat{R}(i,j) = \mathscr{N}_{T,N}(i,j) / \mathscr{N}^*_{T,N}(i) \; .$$

Consequently, one obtains:

$$\widehat{\mathscr{P}}(i,j,t) = \widehat{R}(i,j) \cdot \widehat{F}(i,t) \; .$$

Asymptotic properties (convergence, normality) of these nonparametric estimators are reviewed by Karr (1986), Theorem 8.33. Non-parametric estimation of

the kernel $\mathscr{P}$ of partially observed renewal processes has been considered by Gill (1980) and surveyed by Karr (1986), pp. 347–351.

17.7 Statistical Analysis of Counting Processes

In this section, we present both the statistical analysis of counting processes based on martingale estimators and the extension to these processes of the semiparametric inference initially proposed by Cox (1972, 1975), for duration models. For that purpose, we consider a multivariate counting process with covariates, but our presentation is restricted to the case of a non-censored independent sampling scheme for a counting process derived from a point process.

Let us denote n the number of individuals and i an element of $\{1,\ldots,n\}$. For any individual i, we observe both the path X_t^i of a point process valued in a discrete state space E with K elements and the path of a (multivariate) covariate process $Z^i = (Z_t^i)_t$. These two processes are observed over an interval $[0,T]$ for any i. Given $(Z_t^i)_t$, the $(X_t^i)_t$'s are assumed to be independent. The distribution of $(X_t^i)_t$ is also assumed to be independent of the $(Z_t^j)_t$ for any $j \neq i$, i.e. it is independent of the covariate processes of other individuals.

Now we have to describe the distribution of $(X_t^i)_t$ given $(Z_t^i)_t$. This goal is achieved by representing the point process $(X_t^i)_t$ through a multivariate counting process $(N_t^{i,j})_t$, where $j = (k,\ell), k,\ell \in E,\ k \neq \ell,\ j \in \{1,\ldots,J = K(K-1)\}$. This counting process increases by jumps of size one when the individual i moves from state k to state ℓ. The distribution of $(N_t^{i,j})_t$, given $(Z_t^i)_t$, is characterized by its stochastic intensity with respect to the filtration generated by both the past of *all* the $N_t^{i,j}$ processes, for $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,J\}$, and by *the whole trajectories* of all the (Z_t^i) processes. These stochastic intensities are assumed to take the following form:

$$h_t^{i,j} = \psi^j\left((Z_t^i)_t, \theta\right) \lambda_t^j \, Y_t^{i,j}, i = 1,\ldots,n, \quad j = 1,\ldots,J$$

where:

(i) $\psi^j\left((Z_t^i)_t, \theta\right)$ is a *known positive function* depending on an unknown parameter $\theta \in \Theta \subset \mathbb{R}^p$; in practice, each ψ^j may depend on a subvector of θ only, and then one has to check if the vector θ is identified by the vector of the ψ^j functions;

(ii) λ_t^j is the *baseline intensity function* of $N_t^{i,j}$; it does not depend on individual i; the model is then a proportional hazard type model in which covariates act multiplicatively through the ψ^j functions on the baseline intensity; moreover, λ_t^j is assumed to be non-stochastic and hence a function valued in $\mathbb{R}^+$; in the semiparametric approach, the function λ_t^j is assumed to be totally unknown and the distribution of the X_t^i processes are then characterized by parameters θ and by functions $(\lambda_t^j),\ j = 1,\ldots,J$;

(iii) finally, $Y_t^{i,j}$ is a family of *observable predictable stochastic processes* valued in $\{0,1\}$; we restrict our attention to the case where $Y_t^{i,j}$ characterize the fact that the individual i is "at risk" at time t for jumping from state k to state ℓ, if $j=(k,\ell)$, or equivalently:

$$Y_t^{i,j}=1 \quad \text{if } x_{t^-}^i=k$$
$$Y_t^{i,j}=0 \quad \text{elsewhere}$$

As $h_t^{i,j}$ is the stochastic intensity with respect to all the trajectories of the covariate processes, it would be dependent of all the $(Z_t^i)_t, i=1,\ldots,n$. However, we have assumed that $(X_t^i)_t$ is independent of $(Z_t^{i'})_t$ for any $i'\neq i$ given $(Z_t^i)_t$, and this assumption is expressed by the fact that ψ^j depends only on $(Z_t^i)_t$. In fact, this requirement is not an assumption but is a condition on the definition of the (Z_t^i) processes which may have some elements in common. Moreover, ψ^j may be a function of the whole trajectory of $(Z_t^i)_t$ or of the current value Z_t^i only. The first case requires the continuous-time observation of covariates, which is unrealistic, or some approximation procedures such as discretization of stochastic integrals. The more common case is the one where the instantaneous probability of a jump from state k to state ℓ for the individual i depends only on the current value of the process Z^i, which implies that $\psi^j\big((Z_t^i)_t,\theta\big)$ may be written as $\psi^j(Z_t^i,\theta)$. For example, if (Z_t^i) is a q-dimensional process, a usual specification is the following:

$$\psi^j(Z_t^i,\theta)=\exp(Z_t^{i'}\theta^j)$$

where $\theta^j\in\mathbb{R}^q$ and $\theta=(\theta^j)_{j=1,\ldots,J}$. More generally, such a specification may be constrained by imposing that some components of Z_t^i in the ψ^j function are eliminated.

Up to an additive constant, the log-likelihood of the model is equal to

$$\begin{aligned}
&L_T(\theta,\lambda^1,\ldots,\lambda^J)\\
&\quad=\sum_{i=1}^n\sum_{j=1}^J\int_0^T \ln\,\psi^j(Z_t^i,\theta)\,\mathrm{d}N_t^{i,j}+\sum_{j=1}^J\int_0^T \ln\,\lambda_t^j\,\mathrm{d}\overline{N}_t^j\\
&\quad-\sum_{i=1}^n\sum_{j=1}^J\int_0^T \psi^j(Z_t^i,\theta)\lambda_t^j Y_t^{i,j}\,\mathrm{d}t
\end{aligned}$$

where $\overline{N}_t^j=\sum_{i=1}^n N_t^{i,j}$. The maximum likelihood estimator of θ can be derived from this last equation if the λ_t^j are known functions of unknown parameters. However, the log-likelihood is unbounded if the functions λ_t^j are taken as arguments: λ_t^j may be chosen arbitrarily large at observed jump times (and then the second element in the right hand side of the log-likelihood equation may be as large as desired) and null at other times (and then the third element in the r.h.s. of the log-likelihood equation becomes equal to zero). Then it appears clearly that estimation must be

based on a different procedure: parameters θ can be estimated by maximizing Cox's partial likelihood, and integrals of the $\lambda_t{}^j$'s are then estimated through martingale techniques.

17.7.1 The Cox Likelihood

Following an argument given by Karr (1986), Chap. 5, the Cox's likelihood can be derived as the difference between the log-likelihood function of the observations and the log-likelihood function of the $\overline{N}_t^j$ processes. This difference is a function of θ only, and can be heuristically interpreted as the logarithm of the density of the $N_t^{i,j}$ given $\overline{N}_t^j$ and the covariates. Given the same filtration, intensities of the $\overline{N}_t^j$ processes are the sum over i of $h_t^{i,j}$, i.e.

$$\overline{h}_t^j = \lambda_t{}^j \left(\sum_{i=1}^{n} \psi^j(Z_t^i, \theta) Y_t^{i,j} \right)$$

and the log-likelihood of the statistic $\overline{N}_t^j$ is equal to

$$\begin{aligned} & L_T^*(\theta, \lambda^1, \ldots, \lambda^J) \\ & = \sum_{j=1}^{J} \int_0^T \ln \sum_{i=1}^{n} \left(\psi^j(Z_t^i, \theta) Y_t^{i,j} \right) \mathrm{d}\overline{N}_t^j + \sum_{j=1}^{J} \int_0^T \ln \lambda_t{}^j \, \mathrm{d}\overline{N}_t^j \\ & - \sum_{j=1}^{J} \int_0^T \lambda_t^j \Big(\sum_{i=1}^{n} \psi^j(Z_t^i, \theta) Y_t^{i,j} \Big) \, \mathrm{d}t \end{aligned}$$

The Cox likelihood is then defined as:

$$\begin{aligned} C_T(\theta) &= L_T(\theta, \lambda^1, \ldots, \lambda^J) - L_T^*(\theta, \lambda^1, \ldots, \lambda^J) \\ &= \sum_{j=1}^{J} C_T^j(\theta) \end{aligned}$$

where

$$C_T^j(\theta) = \sum_{i=1}^{n} \int_0^T \ln \, \psi^j(Z_t^i, \theta) \, \mathrm{d}N_t^{i,j} - \int_0^T \ln \Big(\sum_{i=1}^{n} \psi^j(Z_t^i, \theta) Y_t^{i,j} \Big) \, \mathrm{d}\overline{N}_t^j$$

or equivalently:

$$\exp[C_T^j(\theta)] = \frac{\prod_{i=1}^{n} \prod_{\tau_u^{i,j} \le T} \psi^j(Z^i_{\tau_u^{i,j}}, \theta)}{\prod_{\overline{\tau}_u^j \le T} \sum_{i=1}^{n} \psi^j(Z^i_{\overline{\tau}_u^j}, \theta) Y^{i,j}_{\overline{\tau}_u^j}}$$

In this last expression, the second product of the numerator is computed for all the observed jump times $\tau_u^{i,j}$ of the process $N_t^{i,j}$ and the product in the denominator

is computed for all the jump times $\overline{\tau}_u^j$ of the process $\overline{N}_t^j$, i.e. for all the transitions from state k to state ℓ (if $j = (k,\ell)$) observed over all the individuals. Parameters θ are estimated via the maximization of $C_T(\theta)$. Moreover, if $\psi^j(Z_t^i,\theta)$ depends on a subvector θ^j such that all the θ^j's are variation free, the estimator of θ^j may be obtained through a maximization of $C_T^j(\theta) = C_T^j(\theta^j)$ only. In this case, observations of $N_t^{i,j}$ for any i and t are sufficient for the estimation of θ^j.

Asymptotic properties of the maximand of $C_T(\theta)$, denoted $\widehat{\theta}$, have been studied initially by Andersen and Gill (1982) and surveyed, for example, by Karr (1986), Chap. 5. Under usual regularity conditions, it could be shown that $\widehat{\theta}$ is a consistent estimator of θ when n tends to ∞ and that $\sqrt{n}(\widehat{\theta} - \theta)$ is asymptotically normal with variance explicitly given, for example, by Karr (1986), Chap. 5, formulas (5.90a) to (5.91).

17.7.2 The Martingale Estimation of the Integrated Baseline Intensity

For simplicity, let us first present martingale estimators for i.i.d. counting processes, i.e. without presence of covariates. The likelihood of such a model is obtained by setting $\psi^j(Z_t^i,\theta)$ equal to 1 in the log-likelihood function:

$$\begin{aligned} &L_T(\theta,\lambda^1,\ldots,\lambda^J) \\ &\quad = \sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^T \ln\,\psi^j(Z_t^i,\theta)\,\mathrm{d}N_t^{i,j} + \sum_{j=1}^{J}\int_0^T \ln\,\lambda_t^{\,j}\,\mathrm{d}\overline{N}_t^j \\ &\quad - \sum_{i=1}^{n}\sum_{j=1}^{J}\int_0^T \psi^j(Z_t^i,\theta)\lambda_t^{\,j}Y_t^{i,j}\,\mathrm{d}t \end{aligned}$$

In this case, one can easily verify that the log-likelihood is a function of $\overline{N}_t^j$ only, up to an additive constant. This means that these processes constitute a sufficient statistic. Indeed, in this case, the log-likelihood function becomes:

$$L_T(\lambda^1,\ldots,\lambda^J) = \sum_{j=1}^{J}\left[\int_0^T \ln\,\lambda_t^{\,j}\,\mathrm{d}\overline{N}_t^j - \int_0^T \lambda_t^{\,j}\overline{Y}_t^j\,\mathrm{d}t\right]$$

and the processes $\overline{N}_t^j$ have the following stochastic intensities:

$$\overline{h}_t^j = \lambda_t^{\,j}\cdot\overline{Y}_t^j$$

where $\overline{Y}_t^j = \sum_{i=1}^n Y_t^{i,j}$ is the number of individuals at risk for the transition of type j (from state k to state ℓ) at time t.

We want to estimate the integrals of $\lambda_t^{\,j}$ for any j. However, in practice, information is only available for the time interval in which there exists some individuals

from the sample who are at risk for the analyzed transition. Functions of interest are then:

$$\Lambda_t^j = \int_0^t \lambda_s^j \mathbb{I}(\overline{Y}_s^j > 0)\, \mathrm{d}s$$

where $\mathbb{I}(\overline{Y}_s^j > 0) = 1$ if $\overline{Y}_s^j > 0$ and 0 elsewhere.

The *martingale estimator* of Λ_t^j is defined by:

$$\widehat{\Lambda}_t^j = \int_0^t (\overline{Y}_s^j)^{-1} \mathbb{I}(\overline{Y}_s^j > 0)\ \mathrm{d}\overline{N}_s^j$$

This estimator may be heuristically justified by the following argument. Let us start with the differential representation of a counting process:

$$\mathrm{d}\overline{N}_s^j = h_s^j\ \mathrm{d}s + \mathrm{d}M_s^j$$

where M_s^j is a martingale. In our model, this expression becomes

$$\mathrm{d}\overline{N}_s^j = \lambda_s^j \cdot \overline{Y}_s^j \mathrm{d}s + \mathrm{d}M_s^j$$

which can be pre-multiplied by $(\overline{Y}_s^j)^{-1}\ \mathbb{I}(\overline{Y}_s^j > 0)$ to give:

$$(\overline{Y}_s^j)^{-1} \mathbb{I}(\overline{Y}_s^j > 0)\, \mathrm{d}\overline{N}_s^j$$

$$= \lambda_s^j \mathbb{I}(\overline{Y}_s^j > 0)\mathrm{d}s + (\overline{Y}_s^j)^{-1} \mathbb{I}(\overline{Y}_s^j > 0)\ \mathrm{d}M_s^j$$

Integrating the two sides of this relation yields:

$$\widehat{\Lambda}_t^j = \Lambda_t^j + \int_0^t (\overline{Y}_s^j)^{-1}\, \mathbb{I}(\overline{Y}_s^j > 0)\ \mathrm{d}M_s^j$$

The difference between $\widehat{\Lambda}_t^j$ and Λ_t^j is then a stochastic integral of a predictable process with respect to a martingale; so it is a martingale (see Dellacherie and Meyer (1980), Chap. 7, Theorem 3). Moreover, it can be verified that

$$E(\widehat{\Lambda}_t^j - \Lambda_t^j) = 0$$

and

$$< \widehat{\Lambda}_t^j - \Lambda_t^j > = \int_0^t \lambda_s^j (\overline{Y}_s^j)^{-1}\, \mathbb{I}(\overline{Y}_s^j > 0)\, \mathrm{d}s$$

Let us recall that the predictable variation $E_t^j = < \widehat{\Lambda}_t^j - \Lambda_t^j >$ plays the role of an instantaneous variance. In particular:

$$V(\widehat{\Lambda}_t^j - \Lambda_t^j) = E(< \widehat{\Lambda}_t^j - \Lambda_t^j >)$$

Using a martingale estimation approach, E_t^j may be estimated by

$$\widehat{E}_t^j = \int_0^t (\overline{Y}_s^j)^{-2}\, \mathbb{I}(\overline{Y}_s^j > 0)\, \mathrm{d}s$$

Under standard regularity conditions, estimators $\widehat{\Lambda}_t^j$ are asymptotically well-behaved. They are consistent in a strong sense

$$E\left[\sup_t (\widehat{\Lambda}_t^j - \Lambda_t^j)^2\right] \to 0 \text{ when } n \to \infty$$

and $n^{1/2}(\widehat{\Lambda}_t^j - \Lambda_t^j)$ is asymptotically distributed as a centered Gaussian martingale with continuous trajectories and whose predictable variation may be estimated by $n\widehat{E}_t^j$.

Let us now return to the general model with covariates. The differential representation of a process $N_t^{i,j}$ is then:

$$\mathrm{d}N_s^{i,j} = \psi^j(Z_s^i, \theta) \cdot \lambda_s^j \cdot Y_s^{i,j} \mathrm{d}s + \mathrm{d}E_s^{i,j}$$

where $E_s^{i,j}$ is a zero-mean martingale. From the definition of $\mathrm{d}\overline{N}_s^j$, we obtain:

$$d\overline{N}_s^j = \left[\sum_{i=1}^n \psi^j(Z_s^i, \theta) \cdot Y_s^{i,j}\right] \lambda_s^j \mathrm{d}s + \mathrm{d}\overline{E}_s^j$$

in which $\overline{E}_s^j = \sum_{i=1}^n E_s^{i,j}$ is still a zero-mean martingale.
Now let us define:

$$w_t^j(\theta) = \sum_{i=1}^n \psi^j(Z_s^i, \theta) \cdot Y_s^{i,j}$$

Assuming that θ is known, the problem of inference on the integral of the λ_t^j's is identical to the previous case without covariates. The function parameters are now:

$$\Lambda_t^j(\theta) = \int_0^t \mathbb{I}(w_s^j(\theta) > 0)\, \lambda_s^j\, \mathrm{d}s$$

and their estimators are given by:

$$\widehat{\Lambda}_t^j(\theta) = \int_0^t [w_s^j(\theta)]^{-1} \mathbb{I}(w_s^j(\theta) > 0)\, \mathrm{d}\overline{N}_s^j$$

If a Cox procedure is initially used and provides an estimator $\widehat{\theta}$ of θ regardless of the λ_t^j's, an estimator of $\Lambda_t^j(\theta)$ is obtained by substituting $\widehat{\theta}$ for θ in this last expression. It can be proved (see Andersen and Gill (1982)) that asymptotic properties of $\widehat{\Lambda}_t^j(\widehat{\theta})$ are identical to those of $\widehat{\Lambda}_t^j(\theta)$ and that estimators $\widehat{\Lambda}_t^j(\widehat{\theta})$ are independent of each other and independent of $\widehat{\theta}$ asymptotically.

17.8 Conclusions

This chapter focused on definitions and statistical analysis of duration models and point processes. More extensive presentations are contained in textbooks by Kalbfleisch and Prentice (1980), Lawless (1982), Jacobsen (1982), Cox and Oakes (1984), Karr (1986), Daley and Vere-Jones (1988), Lancaster (1990), Andersen, Borgan, Gill and Keiding (1993), or in detailed surveys by Heckman and Singer (1984a), Kiefer (1988), Serfozo (1990), Van den Berg (2001). Markov chains have been completely studied by Chung (1967), Freedman (1971), Revuz (1975) and by Ethier and Kurtz (1986).

References

Abbring, J. and Van den Berg, G. (2003): The nonparametric identification of treatment effects in duration models, *Econometrica*, **71**, 1491–1518.

Andersen, P.K. and Gill, R. (1982): Cox's regression model for counting processes: a large sample study, *Annals of Statistics*, **10**, 1100–1120.

Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993): *Statistical models based on counting processes*. New-York: Springer.

Basawa, I.V. and Prakasa Rao, B.L.S. (1980): *Statistical inference of stochastic processes*. New-York: Academic Press.

Bhattacharya, R.N. and Waymire, E.C. (1990): *Stochastic processes with applications*. New-York: Wiley and Sons.

Billingsley, P. (1961): *Statistical inference for Markov processes*. Chicago: The University of Chicago Press.

Chung, K.L. (1967): *Markov chains with stationary transition probabilities*. New-York: Springer-Verlag.

Cox, D.R. (1972): Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B,* **34**, 187–220.

Cox, D.R. (1975): Partial likelihood, *Biometrika,* **62**, 269–276.

Cox, D.R. and Miller, H.D. (1966): *The theory of stochastic processes*. London: Methuen.

Cox, D.R. and Oakes, D. (1984): *Analysis of survival data*. London: Chapman and Hall.

Daley, D.J. and Vere-Jones, D. (1988): *An introduction to the theory of point processes*. New-York: Springer-Verlag.

Dellacherie, C. and Meyer, P.A. (1980): *Probabilité et potentiel* (Chapitres V à VIII: Théorie des martingales). Paris: Hermann.

Doob, J.L. (1953): *Stochastic processes*. New-York: Wiley and Sons.

Elbers, C. and Ridder, G. (1982): True and spurious dependence: the identifiability of the proportional hazard model, *Review of Economic Studies,* **49**, 403–409.

Ethier, S.N. and Kurtz, T.G. (1986): *Markov processes: characterization and convergence*. New-York: Wiley and Sons

Fougère, D., Le Bihan, H. and Sevestre, P. (2007): Heterogenity in consumer price stickiness: a microeconometric approach, *Journal of Business and Economic Statistics*, **25**, 247–264.

Fourgeaud, C., Gouriéroux, C. and Pradel, J. (1990): Heterogeneity and hazard dominance in duration data models (in french), *Annales d'Economie et de Statistique,* **18**, 1–24.

Freedman, D. (1971): *Markov chains*. San Francisco: Holden-Day.

Gill, R.D. (1980): Non-parametric estimation based on censored observations of a Markov renewal process, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **53**, 97–116.

Han, A. and Hausman, J.A. (1990): Specification and semiparametric estimation of duration models, *Journal of Applied Econometrics,* **5**, 1–28.

Heckman, J.J. and Honoré, B. (1989): The identifiability of the competing risks model, *Biometrika,* **76**, 325–330.

Heckman, J.J. and Singer, B. (1984a): Econometric duration analysis, *Journal of Econometrics,* **24**, 63–132.

Heckman, J.J. and Singer, B. (1984b): The identifiability of the proportional hazards model, *Review of Economic Studies,* **51**, 231–243.

Honoré, B. (1993): Identification results for duration models with multiple spells, *Review of Economic Studies,* **60**, 241–246.

Jacobsen, M. (1982): *Statistical analysis of counting processes*. Berlin: Springer-Verlag.

Kalbfleisch, J.D., and Prentice, R.L. (1980): *The statistical analysis of failure time data*. New York: Wiley.

Karr, A.F. (1986): *Point processes and their statistical inference*. New-York: Marcel Dekker.

Kiefer, N. (1988): Economic duration data and hazard functions, *Journal of Economic Literature,* **XXVI**, 646–679.

Kortram, R.A., Lenstra, A.J., Ridder, G. and Van Rooij A.C.M. (1995): Constructive identification of the mixed proportional hazards model, *Statistica Neerlandica*, **49**, 269–281.

Lancaster, T. (1990): *The Econometric analysis of transition data*, Econometric Society Monographs, Cambridge University Press: New York.

Lawless, J.F. (1982): *Statistical models and methods for lifetime data*. New York: Wiley.

Marshall, A.W. and Olkin, I. (1967): A multivariate exponential distribution, *Journal of the American Statistical Association,* **62**, 30–44.

Mátyás, L. and Sevestre, P. (editors)(1996): *The econometrics of panel data*, Second Revised Edition. Dordrecht: Kluwer Academic Publishers.

Melino A. and Sueyoshi, G.T. (1990): A simple approach to the identifiability of the proportional hazard model, *Economics Letters*, **33**, 63–68.

Mouchart, M. (2004), The econometrics of duration data and of point processes, Lecture Notes, http://www.stat.ucl.ac.be/ISpersonnel/mouchart/DuDa040908$_b$.pdf

Mouchart, M. and J.-M. Rolin (2002): Competing risks models: problems of modelling and of identification, in *Life tables, modelling survival and death,* edited by G. Wunsch, M. Mouchart and J. Duchêne, Dordrecht: Kluwer Academic Publishers, 245–267.

Ridder G. (1990): The nonparametric identification of generalized accelerated failure-time models, *Review of Economic Studies,* **57**, 167–181.

Revuz, D. (1975): *Markov chains*. New-York: North Holland/American Elsevier.

Serfozo, R.F. (1990): Point processes, in *Handbook in operations research and management science*, Vol. 2, edited by D.P. Heyman and M.J. Sobel, Amsterdam: North-Holland, 1–94.

Tsiatis, A.A. (1981): A large sample study of Cox's regression model, *Annals of Statistics,* **9**, 93–108.

Van den Berg G. (2001): Duration models: specification, identification and multiple durations, in *Handbook of econometrics*, Vol. 5, edited by J. Heckman and E.Leamer, Amsterdam: North-Holland, 3381–3460.

Chapter 18
GMM for Panel Data Count Models

Frank Windmeijer

18.1 Introduction

This chapter gives an account of the recent literature on estimating (dynamic) models for panel count data. Specifically, the treatment of unobserved individual heterogeneity that is correlated with the explanatory variables and the presence of explanatory variables that are not strictly exogenous are central. Moment conditions are discussed for these types of problems that enable estimation of the parameters by the Generalised Method of Moments (GMM). Interest in exponential regression models has increased substantially in recent years. The Poisson regression model for modelling an integer count dependent variable is an obvious example where the conditional mean function is routinely modelled to be exponential. But also models for continuous positive dependent variables that have a skewed distribution are increasingly being advocated to have an exponential conditional mean function. Although for these data the log transformation can be applied to multiplicative models, the "retransformation" problem often poses severe difficulties if the object of interest is the level of for example costs, see e.g. Manning, Basu and Mullahy (2005). Santos Silva and Tenreyro (2006) also strongly recommend to estimate the multiplicative models directly, as the log transformation can be unduly restrictive. Although the focus of this chapter is on models for count data, almost all procedures can directly be applied to models where the dependent variable is a positive continuous variable and the conditional mean function is exponential. The one exception is the linear feedback model as described in Sect. 18.3.4, which is a dynamic model specification specific to discrete count data.

Section 18.2 discusses instrumental variables estimation for count data models in cross sections. Section 18.3 derives moment conditions for the estimation of (dynamic) models for count panel data allowing for correlated fixed effects and

Frank Windmeijer
Department of Economics, University of Bristol, 8 Woodland Road, Bristol BS8 1TN, UK, e-mail: f.windmeijer@bristol.ac.uk

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

weakly exogenous regressors. Section 18.4 discusses GMM estimation. Section 18.5 reviews some of the applied literature and software to estimate the models by nonlinear GMM. As standard Wald tests based on efficient two-step GMM estimation results are known to have poor finite sample behavior, Sect. 18.6 considers alternative test procedures that have recently been proposed in the literature. It also considers estimation by the continuous updating estimator (CUE) as this estimator has been shown to have a smaller finite sample bias than one- and two-step GMM. As asymptotic standard errors for the CUE are downward biased in finite samples we use results from alternative, many weak instrument asymptotics that lead to a larger asymptotic variance of the CUE.

18.2 GMM in Cross-Sections

The Poisson distribution for an integer count variable y_i, $i = 1,\ldots,N$, with mean μ_i is given by

$$P(y_i) = \frac{e^{-\mu_i}\mu^{y_i}}{y_i!}$$

and the Poisson regression model specifies $\mu_i = \exp(x_i'\beta)$, where x_i is a vector of explanatory variables and β a parameter vector to be estimated. The log-likelihood function for the sample is then given by

$$\ln L = \sum_{i=1}^{N} y_i \ln(\mu_i) - \mu_i - \ln(y_i!)$$

with first-order condition

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{N} x_i (y_i - \mu_i) = 0\,. \tag{18.1}$$

It is therefore clear that the Poisson regression estimator is a method of moments estimator. If we write the model with an additive error term u_i as

$$y_i = \exp\left(x_i'\beta\right) + u_i = \mu_i + u_i$$

with

$$E(x_i u_i) = E(x_i(y_i - \mu_i)) = 0\,,$$

this is clearly the population equivalent of the sample first order condition in the Poisson regression model.

An alternative moment estimator is obtained by specifying the error term as multiplicative in the model

$$y_i = \exp\left(x_i'\beta\right) w_i = \mu_i w_i$$

with associated moment conditions

$$E\left(\left(w_i-1\right)|x_i\right)=E\left(\left(\frac{y_i-\mu_i}{\mu_i}\right)|x_i\right)=0. \tag{18.2}$$

Mullahy (1997) was the first to introduce GMM instrumental variables estimation of count data models with endogenous explanatory variables. He used the multiplicative setup with x_i being correlated with the unobservables w_i such that $E\left(\left(w_i-1\right)|x_i\right)\neq 0$ and the moment estimator that solves (18.2) is therefore not consistent. There are instruments z_i available that are correlated with the endogenous regressors, but not with w_i such that

$$E\left(\left(w_i-1\right)|z_i\right)=E\left(\left(\frac{y_i-\mu_i}{\mu_i}\right)|z_i\right)=0. \tag{18.3}$$

Denote[1]

$$g_i=z_i\left(\frac{y_i-\mu_i}{\mu_i}\right),$$

then the GMM estimator for β that minimises

$$Q_N\left(\beta\right)=\left(\frac{1}{N}\sum_{i=1}^{N}g_i\right)'W_N^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}g_i\right)$$

is consistent, where W_N is a weight matrix. The efficient two-step weight matrix is given by

$$W_N\left(\widehat{\beta}_1\right)=\frac{1}{N}\sum_{i=1}^{N}g_i\left(\widehat{\beta}_1\right)g_i\left(\widehat{\beta}_1\right)'$$

where

$$g_i\left(\widehat{\beta}_1\right)=z_i\left(\frac{y_i-\exp\left(x_i'\widehat{\beta}_1\right)}{\exp\left(x_i'\widehat{\beta}_1\right)}\right)$$

with $\widehat{\beta}_1$ an initial consistent estimator. Angrist (2001) strengthens the arguments for using these moment conditions for causal inference as he shows that in a model with endogenous treatment and a binary instrument, the Mullahy procedure estimates a proportional local average treatment effect (LATE) parameter in models with no covariates.

Windmeijer and Santos Silva (1997) propose use of the additive moment conditions

$$E\left(\left(y_i-\mu_i\right)|z_i\right)=0, \tag{18.4}$$

estimating the parameters β again by GMM, with in this case $g_i=z_i\left(y_i-\mu_i\right)$. They and Mullahy (1997) compare the two sets of moment conditions and show that both sets cannot in general be valid at the same time. One exception is when there is

[1] From the conditional moments (18.3) it follows that any function $h\left(z\right)$ are valid instruments, which raises the issue of optimal instruments. Here, we will only consider $h\left(z\right)=z$.

classical measurement error in an explanatory variable, as in that case both additive and multiplicative moment conditions are valid. Consider the simple model

$$y_i = \exp(\alpha + x_i^* \beta) + u_i$$

but x_i^* is not observed. Instead we observe x_i

$$x_i = x_i^* + \varepsilon_i$$

and estimate β in the model

$$y_i = \exp(\alpha + x_i\beta - \varepsilon_i\beta) + u_i \, .$$

When instruments z_i are available that are correlated with x_i and independent of the i.i.d measurement errors ε_i, then the multiplicative moment conditions

$$E\left(\left(\frac{y_i - \widetilde{\mu}_i}{\widetilde{\mu}_i}\right)|z_i\right) = 0$$

are valid, where

$$\begin{aligned} \widetilde{\mu}_i &= \exp(\widetilde{\alpha} + x_i\beta) \\ \widetilde{\alpha} &= \alpha + \ln(E[\exp(-\varepsilon\beta)]) \, , \end{aligned}$$

and the latter expectation is assumed to be a constant. However, also the additive moment conditions are valid as

$$E((y_i - \widetilde{\mu}_i)|z_i) = 0.$$

18.3 Panel Data Models

Let y_{it} denote the discrete count variable to be explained for subject i, $i = 1, \ldots, N$, at time t, $t = 1, \ldots, T$; and let x_{it} denote a vector of explanatory variables. An important feature in panel data applications is unobserved heterogeneity or individual fixed effects. For count data models these effects are generally modelled multiplicatively as

$$\begin{aligned} y_{it} &= \exp\left(x_{it}'\beta + \eta_i\right) + u_{it} \\ &= \mu_{it}\nu_i + u_{it} \, , \end{aligned}$$

where $\nu_i = \exp(\eta_i)$ is a permanent scaling factor for the individual specific mean. In general, it is likely that the unobserved heterogeneity components η_i are correlated with the explanatory variables, $E(x_{it}\eta_i) \neq 0$, and therefore standard random effects estimators for β will be inconsistent, see Hausman, Hall and Griliches (1984). This section will describe moment conditions that can be used to consistently estimate the parameters β when there is correlation between η_i and x_{it} and allowing for different

exogeneity properties of the explanatory variables, i.e. the regressors being strictly exogenous, predetermined or endogenous. Throughout we assume that the u_{it} are not serially correlated and that $E(u_{it}|\nu_i) = 0, t = 1,\ldots,T$.

18.3.1 Strictly Exogenous Regressors

When the x_{it} are *strictly exogenous,* there is no correlation between any of the idiosyncratic shocks $u_{is}, s = 1,\ldots,T$ and any of the $x_{it}, t = 1,\ldots T$, and the conditional mean of y_{it} satisfies

$$E(y_{it}|\nu_i, x_{it}) = E(y_{it}|\nu_i, x_{i1},\ldots,x_{iT}) \ .$$

For this case, Hausman, Hall and Griliches (1984) use the Poisson conditional maximum likelihood estimator (CMLE), conditioning on $\sum_{t=1}^{T} y_{it}$, which is the sufficient statistic for η_i. This method mimics the fixed effect logit approach of Chamberlain (1984). However, the Poisson maximum likelihood estimator (MLE) for β in a model with separate individual specific constants does not suffer from the incidental parameters problem, and is therefore consistent and the same as the CMLE. To see this, note that the maximum likelihood first order conditions for the ν_i are given by

$$\frac{\partial \ln L}{\partial \nu_i} = \sum_{t=1}^{T} \frac{\partial (y_{it} \ln(\mu_{it}\nu_i) - \mu_{it}\nu_i)}{\partial \nu_i} = \sum_{t=1}^{T} \left(\frac{y_{it}}{\nu_i} - \mu_{it} \right) = 0$$

and therefore the MLE for ν_i is given by

$$\widehat{\nu}_{i(ML)} = \frac{\overline{y}_i}{\overline{\mu}_i} \ ,$$

where $\overline{y}_i = T^{-1} \sum_{t=1}^{T} y_{it}$ and $\overline{\mu}_i = T^{-1} \sum_{t=1}^{T} \exp(x_{it}'\beta)$. The MLE of the fixed effect is independent of ν_i. Substituting the fixed effects estimates in the first order conditions for β results in the moment conditions

$$\frac{\partial \ln L}{\partial \beta}(\widehat{\nu}_i) = \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - \mu_{it} \frac{\overline{y}_i}{\overline{\mu}_i} \right) x_{it} = 0 \ .$$

When x_{it} is strictly exogenous,

$$\text{plim}_{N\to\infty} \frac{1}{N} \frac{\partial \ln L}{\partial \beta}(\widehat{\nu}_i) = \text{plim}_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(u_{it} - \frac{\mu_{it}}{\overline{\mu}_i} \overline{u}_i \right) x_{it} = 0 \ ,$$

with $\overline{u}_i = T^{-1} \sum_{t=1}^{T} u_{it}$, and therefore the MLE for β is consistent.[2] It is further identical to the CMLE. The latter can be seen as follows. The Poisson conditional log likelihood function is given by

[2] Lancaster (2002) finds the same result for the Poisson model by means of a decomposition of the likelihood.

$$\ln CL = \sum_{i=1}^{N}\sum_{t=1}^{T}\Gamma(y_{it}+1) - \sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}\ln\left[\sum_{s=1}^{T}\exp\left(-(x_{it}-x_{is})'\beta\right)\right],$$

where $\Gamma(.)$ is the gamma function, see Hausman, Hall and Griliches (1984, p. 919). The first-order condition for β is

$$\begin{aligned}\frac{\partial \ln CL}{\partial \beta} &= \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{y_{it}}{\sum_{s=1}^{T}\exp\left(-(x_{it}-x_{is})'\beta\right)}\sum_{s=1}^{T}\exp\left(-(x_{it}-x_{is})'\beta\right)(x_{it}-x_{is}) \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}x_{it} - \sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}\frac{\sum_{s=1}^{T} x_{is}\exp(x_{is}'\beta)}{\sum_{s=1}^{T}\exp(x_{is}'\beta)} \\ &= \sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}\left(y_{it} - \mu_{it}\frac{\overline{y}_i}{\overline{\mu}_i}\right),\end{aligned}$$

which is exactly the same as the MLE first order condition for β.

The first order conditions imply that the Poisson MLE for β is equivalent to the moment estimator in a model where the ratio of individual, or within group, means are used to approximate the individual specific effects. This *mean scaling* model is given by

$$y_{it} = \mu_{it}\frac{\overline{y}_i}{\overline{\mu}_i} + u_{it}^{*}, \tag{18.5}$$

where $u_{it}^{*} = u_{it} - \frac{\mu_{it}}{\overline{\mu}_i}\overline{u}_i$. Blundell, Griffith and Windmeijer (2002) call the resulting estimator the *within group mean scaling* estimator.[3]

18.3.2 Predetermined Regressors

A regressor is predetermined when it is not correlated with current and future shocks, but it is correlated with past shocks:

$$\begin{aligned}E\left(x_{it}u_{it+j}\right) &= 0, \quad j \geq 0 \\ E\left(x_{it}u_{it-s}\right) &\neq 0, \quad s \geq 1\,.\end{aligned}$$

With predetermined regressors, the within group mean scaling estimator is no longer consistent. Chamberlain (1992) has proposed a transformation that eliminates the fixed effect from the multiplicative model and generates orthogonality conditions that can be used for consistent estimation in count data models with predetermined regressors. The quasi-differencing transformation is

$$s_{it} = y_{it}\frac{\mu_{it-1}}{\mu_{it}} - y_{it-1} = u_{it}\frac{\mu_{it-1}}{\mu_{it}} - u_{it-1}\,.$$

[3] Clearly, the Poisson pseudo-likelihood results are preserved, see also Wooldridge (1999).

Let $x_i^{t-1} = (x_{i1}, \ldots, x_{it-1})$. When x_{it} is predetermined, the following moment conditions hold:

$$E\left(s_{it}|x_i^{t-1}\right) = E\left(\left(E\left(u_{it}|x_i^t\right)\frac{\mu_{it-1}}{\mu_{it}} - u_{it-1}\right)|x_i^{t-1}\right) = 0. \tag{18.6}$$

Wooldridge (1991) proposed the following quasi-differencing transformation

$$q_{it} = \frac{y_{it}}{\mu_{it}} - \frac{y_{it-1}}{\mu_{it-1}} = \frac{u_{it}}{\mu_{it}} - \frac{u_{it-1}}{\mu_{it-1}} ,$$

with moment conditions

$$E\left(q_{it}|x_i^{t-1}\right) = E\left(\left(\frac{E\left(u_{it}|x_i^t\right)}{\mu_{it}} - \frac{u_{it-1}}{\mu_{it-1}}\right)|x_i^{t-1}\right) = 0 .$$

It is clear that a variable in x_{it} can not have only non-positive or non-negative values, as then the corresponding estimate for β is infinity. A way around this problem is to transform x_{it} in deviations from its overall mean, $\widetilde{x}_{it} = x_{it} - \bar{x}$, with $\bar{x} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}$, see Windmeijer (2000).

Both moment conditions can also be derived from a multiplicative model specification

$$y_{it} = \exp\left(x_{it}'\beta + \eta_i\right) w_{it} = \mu_{it}\nu_i w_{it} ,$$

where x_{it} is now predetermined w.r.t. w_{it}. Again, we assume that the w_{it} are not serially correlated and not correlated with ν_i, and $E\left(w_{it}\right) = 1$. The Chamberlain quasi-differencing transformation in this case is equivalent to

$$s_{it} = y_{it}\frac{\mu_{it-1}}{\mu_{it}} - y_{it-1} = \nu_i\mu_{it-1}\left(w_{it} - w_{it-1}\right) ,$$

with moment conditions

$$E\left(s_{it}|x_i^{t-1}\right) = E\left(\nu_i\mu_{it-1}E\left(\left(w_{it} - w_{it-1}\right)|\nu_i, x_i^{t-1}\right)|x_i^{t-1}\right) = 0 .$$

Equivalently, for the Wooldridge transformation,

$$q_{it} = \frac{y_{it}}{\mu_{it}} - \frac{y_{it-1}}{\mu_{it-1}} = \nu_i\left(w_{it} - w_{it-1}\right)$$

and

$$E\left(q_{it}|x_i^{t-1}\right) = E\left(\nu_i E\left(\left(w_{it} - w_{it-1}\right)|\nu_i, x_i^{t-1}\right)|x_i^{t-1}\right) = 0 .$$

18.3.3 Endogenous Regressors

Regressors are endogenous when they are correlated with current (and possibly past) shocks $E\left(x_{it}u_{it-s}\right) \neq 0$, $s \geq 0$, for the specification with additive errors u_{it}, or when

$E(x_{it}w_{it-s}) \neq 0$, $s \geq 0$, for the specification with multiplicative errors w_{it}. It is clear from the derivations in the previous section that we cannot find valid sequential conditional moment conditions for the specification with additive errors due to the non-separability of the u_{it} and μ_{it}. For the multiplicative error specification, there is again non-separability of μ_{it-1} and $(w_{it} - w_{it-1})$ for the Chamberlain transformation and so

$$E\left(s_{it}|x_i^{t-2}\right) = E\left(v_i\mu_{it-1}E\left((w_{it} - w_{it-1})\,|v_i, x_i^{t-1}\right)|x_i^{t-2}\right) \neq 0\,.$$

In contrast, the Wooldridge transformation does not depend on μ_{it} or μ_{it-1} in this case. Valid moment conditions are then given by

$$E\left(q_{it}|x_i^{t-2}\right) = E\left(v_iE\left((w_{it} - w_{it-1})\,|v_i, x_i^{t-2}\right)|x_i^{t-2}\right) = 0\,.$$

Therefore, in the case of endogenous explanatory variables, only the Wooldridge transformation can be used for the consistent estimation of the parameters β. This includes the case of classical measurement error in x_{it}, where the measurement error is not correlated over time.

18.3.4 Dynamic Models

Specifying dynamic models for count data by including lags of the dependent count variables in the explanatory part of the model is not as straightforward as with linear models for a continuous dependent variable. Inclusion of the lagged dependent variable in the exponential mean function may lead to rapidly exploding series. A better starting point is to specify the model as in Crépon and Duguet (1997)

$$y_{it} = h(y_{it-1}, \gamma)\exp\left(x_{it}'\beta + \eta_i\right) + u_{it}\,,$$

where $h(.,.) > 0$ is any given function describing the way past values of the dependent variable are affecting the current value.

Let

$$d_{it} = 1_{\{y_{it}=0\}}\,,$$

then a possible choice for $h(.,.)$ is

$$h(y_{it}, \gamma) = \exp\left(\gamma_1 \ln\left(y_{it-1} + cd_{it-1}\right) + \gamma_2 d_{it-1}\right)\,,$$

where c is a pre-specified constant. In this case, $\ln(y_{it-1})$ is included as a regressor for positive y_{it-1}, and zero values of y_{it-1} have a separate effect on current values of y_{it}. Crépon and Duguet (1997) considered

$$h(y_{it}, \gamma) = \exp\left(\gamma(1 - d_{it-1})\right)\,,$$

and extensions thereof to several regime indicators.

Blundell, Griffith and Windmeijer (2002) propose use of a linear feedback model for modelling dynamic count panel data process. The linear feedback model of order 1 (LFM(1)) is defined as

$$\begin{aligned} y_{it} &= \gamma y_{it-1} + \exp(x_{it}'\beta + \eta_i) + u_{it} \\ &= \gamma y_{it-1} + \mu_{it}\nu_i + u_{it} \ , \end{aligned}$$

where the lag of the dependent variable enters the model linearly. Extending the model to include further lags is straightforward. The LFM has its origins in the Integer-Valued Autoregressive (INAR) process and can be motivated as an entry-exit process with the probability of exit equal to $(1-\gamma)$. The correlation over time for the INAR(1) process without additional regressors is similar to that of the AR(1) model, $\text{corr}\,(y_{it}, y_{it-j}) = \gamma^j$.

For the patents-R&D model, Blundell, Griffith and Windmeijer (2002) consider the economic model

$$P_{it} = k\left(R_{it}^{\beta} + (1-\delta)R_{it-1}^{\beta} + (1-\delta)^2 R_{it-2}^{\beta}\ldots\right)\nu_i + \varepsilon_{it} \tag{18.7}$$

where P_{it} and R_{it} are the number of patents and R&D expenditures for firm i at time t respectively, k is a positive constant and R&D expenditures depreciate geometrically at rate δ. The long run steady state for firm i, ignoring feedback from patents to R&D, can be written as

$$P_i = \frac{k}{\delta} R_i^{\beta}\nu_i \ ,$$

and β can therefore be interpreted as the long run elasticity. Inverting (18.7) leads to

$$P_{it} = kR_i^{\beta}\nu_i + (1-\delta)P_{it-1} + u_{it}$$

and so in the LFM model

$$P_{it} = \gamma P_{it-1} + \exp\left(k^* + \beta \ln(R_{it})\right)\nu_i + u_{it}$$

the estimate for γ is an estimate of the depreciation factor $(1-\delta)$ and the estimate for β is an estimate of the *long run* elasticity of the R&D returns to patents.

Even when the x_{it} are strictly exogenous, the within group mean scaling estimator will be inconsistent for small T, as the lagged dependent variable is a predetermined variable. For estimation by GMM, the Chamberlain quasi-differencing transformation for the LFM(1) model is given by

$$s_{it} = (y_{it} - \gamma y_{it-1})\frac{\mu_{it-1}}{\mu_{it}} - (y_{it-1} - \gamma y_{it-2}) \tag{18.8}$$

and the Wooldridge quasi-differencing transformation is given by

$$q_{it} = \frac{y_{it} - \gamma y_{it-1}}{\mu_{it}} - \frac{y_{it-1} - \gamma y_{it-2}}{\mu_{it-1}} \ .$$

For predetermined x_{it} the following moment conditions hold

$$E(s_{it}|y_i^{t-2},x_i^{t-1}) = 0;$$
$$E(q_{it}|y_i^{t-2},x_i^{t-1}) = 0,$$

while for endogenous x_{it}, only the Wooldridge moment conditions are valid

$$E(q_{it}|y_i^{t-2},x_i^{t-2}) = 0\,.$$

18.4 GMM

The orthogonality conditions as described in the sections above can be used to consistently estimate the model parameters by the GMM estimation technique (see Hansen, 1982). Let θ be the k-vector of parameters to be estimated, e.g. for the LFM(1) model $\theta = (\gamma,\beta')'$. The model has a true parameter θ_0 satisfying the q moment conditions

$$E\left[g_i(\theta_0)\right] = 0\,.$$

The GMM estimator $\widehat{\theta}$ for θ_0 is defined as

$$\widehat{\theta} = \arg\min_{\theta\in\Theta}\left[\frac{1}{N}\sum_{i=1}^{N} g_i(\theta)\right]' W_N^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} g_i(\theta)\right]\,,$$

where Θ is a compact set of parameter values; W_N satisfies $\mathrm{plim}_{N\to\infty} W_N = W$, with W a positive definite matrix. Regularity conditions are assumed such that $\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} g_i(\theta) = E\left[g_i(\theta)\right]$ and $\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g_i(\theta_0) \to N(0,\Psi)$ where $\Psi = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} E\left[\left(g_i(\theta_0)\,g_i(\theta_0)'\right)\right]$. Let $\Gamma(\theta) = E\left[\partial g_i(\theta)/\partial\theta'\right]$ and $\Gamma_{\theta_0}\equiv\Gamma(\theta_0)$, then $\sqrt{N}\left(\widehat{\theta}-\theta_0\right)$ has a limiting normal distribution,

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right) \to N(0,V_W)\,,$$

where

$$V_W = \left(\Gamma_{\theta_0}' W^{-1}\Gamma_{\theta_0}\right)^{-1}\Gamma_{\theta_0}' W^{-1}\Psi W^{-1}\Gamma_{\theta_0}\left(\Gamma_{\theta_0}' W^{-1}\Gamma_{\theta_0}\right)^{-1}\,.$$

The efficient two-step GMM estimator, denoted $\widehat{\theta}_2$, is based on a weight matrix that satisfies $\mathrm{plim}_{N\to\infty} W_N = \Psi$, with $V_W = \left(\Gamma_{\theta_0}'\Psi^{-1}\Gamma_{\theta_0}\right)^{-1}$ in that case. A weight matrix that satisfies this property is given by

$$W_N\left(\widehat{\theta}_1\right) = \frac{1}{N}\sum_{i=1}^{N} g_i\left(\widehat{\theta}_1\right) g_i\left(\widehat{\theta}_1\right)'\,,$$

where $\widehat{\theta}_1$ is an initial consistent estimator for θ_0.

Denote $\overline{g}(\theta) = \frac{1}{N}\sum_{i=1}^{N} g_i(\theta)$. The standard test for overidentifying restrictions is N times the minimised GMM criterion

$$NQ_{W_N}\left(\widehat{\theta}_2\right) = N\overline{g}\left(\widehat{\theta}_2\right)' W_N^{-1}\left(\widehat{\theta}_1\right)\overline{g}\left(\widehat{\theta}_2\right) ,$$

which has an asymptotic chi-squared distribution with $q-k$ degrees of freedom when the moment conditions are valid.

For the Chamberlain quasi-differencing transformation the GMM estimator $\widehat{\theta}$ minimises

$$\widehat{\theta} = \arg\min_{\theta\in\Theta}\left(\frac{1}{N}\sum_{i=1}^{N} s_i(\theta)' Z_i\right) W_N^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i' s_i(\theta)\right) ,$$

where, for the LFM(1) model, $s_i(\theta)$ is the $T-2$ vector $(s_{i3}, s_{i4}, \ldots, s_{iT})'$, with s_{it} as defined in (18.8), Z_i is the matrix of instruments and W_N is a weight matrix. When the full sequential set of instruments is used and x_{it} is predetermined, the instrument matrix for the LFM(1) model is given by

$$Z_i = \begin{bmatrix} y_{i1} & x_{i1} & x_{i2} & & & & & \\ & & & \ddots & & & & \\ & & & & y_{i1} & \cdots & y_{iT-2} & x_{i1} & \cdots & x_{iT-1} \end{bmatrix} .$$

The efficient weight matrix is

$$W_N\left(\widehat{\theta}_1\right) = \frac{1}{N}\sum_{i=1}^{N} Z_i' s_i(\widehat{\theta}_1) s_i(\widehat{\theta}_1)' Z_i ,$$

where $\widehat{\theta}_1$ can be a GMM estimator using for example $W_N = \frac{1}{N}\sum_{i=1}^{N} Z_i' Z_i$ as the initial weight matrix. As stated above, under the assumed regularity conditions both $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are asymptotically normally distributed. The asymptotic variance of $\widehat{\theta}_1$ is computed as

$$\begin{aligned} \widehat{\mathrm{var}}\left(\widehat{\theta}_1\right) = \frac{1}{N}&\left(C\left(\widehat{\theta}_1\right)' W_N^{-1} C\left(\widehat{\theta}_1\right)\right)^{-1} C\left(\widehat{\theta}_1\right)' W_N^{-1} W_N\left(\widehat{\theta}_1\right) W_N^{-1} C\left(\widehat{\theta}_1\right) \\ &\times\left(C\left(\widehat{\theta}_1\right)' W_N^{-1} C\left(\widehat{\theta}_1\right)\right)^{-1} \end{aligned}$$

where

$$C\left(\widehat{\theta}_1\right) = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial Z_i' s_i(\theta)}{\partial\theta}\Big|_{\widehat{\theta}_1} .$$

The asymptotic variance of the efficient two-step GMM, estimator is computed as

$$\widehat{\mathrm{var}}\left(\widehat{\theta}_2\right) = \frac{1}{N}\left(C\left(\widehat{\theta}_2\right)' W_N^{-1}\left(\widehat{\theta}_1\right) C\left(\widehat{\theta}_2\right)\right)^{-1} .$$

18.5 Applications and Software

The instrumental variables methods for count data models with endogenous regressors using cross section data, as described in Sect. 18.2, have often been applied in the health econometrics literature. For example, Mullahy (1997) uses the multiplicative moment conditions to estimate cigarette demand functions with a habit stock measure as endogenous regressor. Windmeijer and Santos Silva (1997) estimate health care demand functions with a self reported health measure as possible endogenous variable, while Vera-Hernandez (1999) and Schellhorn (2001) estimate health care demand functions with endogenous insurance choice. An example outside the health econometrics literature is Kelly (2000) who models the number of crimes with police activity as an endogenous regressor.

The count panel data literature has largely focused on estimating models for patenting and the returns to R&D investments, which started with the seminal paper of Hausman, Hall and Griliches (1984). Following the development of the quasi-differencing approach of Wooldridge (1991, 1997), Chamberlain (1992), and Montalvo (1997), Cincera (1997), Crépon and Duguet (1997), Blundell, Griffith and Van Reenen (1999) and Blundell, Griffith and Windmeijer (2002) developed and/or estimated patent (or innovation) production functions using the GMM framework allowing for correlated firm specific effects and weakly exogenous inputs. More recently, Kim and Marschke (2005) use the GMM framework to find a relationship between a firms' patenting behavior and scientist turnover, whereas Salomon and Shaver (2005) estimate a linear feedback model and find that exporting has a positive effect on innovating behavior of the firm.

The latter two publications estimated the models using *ExpEnd*, Windmeijer (2002). This is a user friendly open source GAUSS (Aptech Systems, 2005) code for nonlinear GMM estimation of the models described in the previous sections.[4] For cross-section data, *ExpEnd* estimates simple robust Poisson regression models using moment conditions (18.1); and instrumental variables regressions using Mullahy's (1997) multiplicative moment conditions (18.3) or the additive moment conditions (18.4). For panel data, *ExpEnd* estimates pooled robust Poisson regression models; fixed effects models, using the mean scaling model (18.5); and the quasi-differencing models using the Chamberlain (1992) or the Wooldridge (1991, 1997) transformation, for static, distributed lag and linear feedback models. For the quasi-differencing models, sequential and so-called stacked IV type instruments can be specified, in both cases allowing for a flexible lag length of the instruments. For overidentified models one- and two-step GMM parameter estimates are reported, together with asymptotic standard errors. The test for overidentifying restrictions is reported and for the panel data models the output further includes a test of first and second order serial correlation of the quasi-differencing "residuals" $s_{it}(\widehat{\theta})$ or $q_{it}(\widehat{\theta})$. If the model is correctly specified one expects to find an MA(1) serial correlation structure.

[4] For a review, see Romeu (2004).

Another package that enables researchers to estimate these model types is TSP Hall and Cummins (2005). Kitazawa (2000) provides various TSP procedures for the estimation of count panel data models. Also LIMDEP, Greene (2005), provides an environment where these models can be estimated.

18.6 Finite Sample Inference

Standard Wald tests based on two-step efficient GMM estimators are known to have poor finite sample properties (see e.g. Blundell and Bond (1998)). Bond and Windmeijer (2005) therefore analysed the finite sample performance of various alternative test procedures for testing linear restrictions in linear panel data models. The statistics they found to perform well in Monte Carlo exercises were an alternative two-step Wald test that uses a finite sample correction for the asymptotic variance matrix, the LM test, and a simple criterion-based test. In this section we briefly describe these procedures and adapt them to the case of nonlinear GMM estimation where necessary.

Newey and Smith (2004) have shown that the GMM estimator can further also suffer from quite large finite sample biases and advocate use of Generalized Empirical Likelihood (GEL) estimators that they show to have smaller finite sample biases. We will consider here the performance of the Continuous Updating Estimator (CUE) as proposed by Hansen, Heaton and Yaron (1996), which is a GEL estimator. The Wald test based on the CUE has also been shown to perform poorly in finite samples by e.g. Hansen, Heaton and Yaron (1996). Newey and Windmeijer (2005) derive the asymptotic distribution of the CUE when there are many weak moment conditions. The asymptotic variance in this case is larger than the usual asymptotic one and we will analyse the performance of an alternative Wald test that uses an estimate for this larger asymptotic variance, together with a criterion based test for the CUE as proposed by Hansen, Heaton and Yaron (1996).

The estimators and test procedures will be evaluated in a Monte Carlo study of testing linear restrictions in a static count panel data model with an explanatory variable that is correlated with the fixed unobserved heterogeneity and which is predetermined. The Chamberlain quasi-differencing transformation will be used with sequential moment conditions.

18.6.1 Wald Test and Finite Sample Variance Correction

The standard Wald test for testing r linear restrictions of the form $r(\theta_0) = 0$ is calculated as

$$\text{Wald} = r\left(\widehat{\theta}\right)' \left(R'\widehat{\text{var}}\left(\widehat{\theta}\right)R\right)^{-1} r\left(\widehat{\theta}\right) ,$$

where $R = \partial r(\theta)/\partial\theta'$, and has an asymptotic χ^2_r distribution under the null. Based on the two-step GMM estimator and using its conventional asymptotic variance estimate, the Wald test has often been found to overreject correct null hypotheses severely compared to its nominal size. This can occur even when the estimator has negligible finite sample bias, due to the fact that the estimated asymptotic standard errors can be severely downward biased in small samples. Windmeijer (2005) proposed a finite sample variance correction that takes account of the extra variation due to the presence of the estimated parameters $\widehat{\theta}_1$ in the weight matrix. He showed by means of a Monte Carlo study that this correction works well for in linear models, but it is not clear how well it will work in nonlinear GMM.

To derive the finite sample corrected variance, let

$$\overline{g}(\theta) = \frac{1}{N}\sum_{i=1}^{N} g_i(\theta);\ C(\theta) = \frac{\partial \overline{g}(\theta)}{\partial\theta'};\ G(\theta) = \frac{\partial C(\theta)}{\partial\theta},$$

and

$$b_{\theta_0,W_N} = \frac{1}{2}\frac{\partial Q_{W_N}}{\partial\theta}|_{\theta_0} = C(\theta_0)' W_N^{-1}\overline{g}(\theta_0);$$
$$A_{\theta_0,W_N} = \frac{1}{2}\frac{\partial^2 Q_{W_N}}{\partial\theta\partial\theta'}|_{\theta_0} = C(\theta_0)' W_N^{-1} C(\theta_0) + G(\theta_0)'\left(I_k \otimes W_N^{-1}\overline{g}(\theta_0)\right).$$

A standard first order Taylor series approximation of $\widehat{\theta}_2$ around θ_0, conditional on $W_N\left(\widehat{\theta}_1\right)$, results in

$$\widehat{\theta}_2 - \theta_0 = -A^{-1}_{\theta_0,W_N\left(\widehat{\theta}_1\right)} b_{\theta_0,W_N\left(\widehat{\theta}_1\right)} + O_p\left(N^{-1}\right).$$

A further expansion of $\widehat{\theta}_1$ around θ_0 results in

$$\widehat{\theta}_2 - \theta_0 = -A^{-1}_{\theta_0,W_N(\theta_0)} b_{\theta_0,W_N(\theta_0)} + D_{\theta_0,W_N(\theta_0)}\left(\widehat{\theta}_1 - \theta_0\right) + O_p\left(N^{-1}\right), \qquad (18.9)$$

where

$$W_N(\theta_0) = \frac{1}{N}\sum_{i=1}^{N} g_i(\theta_0)\, g_i(\theta_0)'$$

and

$$D_{\theta_0,W_N(\theta_0)} = \frac{\partial}{\partial\theta'}\left(-A^{-1}_{\theta_0,W_N(\theta)} b_{\theta_0,W_N(\theta)}\right)|_{\theta_0}$$

is a $k \times k$ matrix.

Let $\widehat{\theta}_1$ be a one-step GMM estimator that uses a weight matrix W_N that does not depend on estimated parameters. An estimate of the variance of $\widehat{\theta}_2$ that incorporates the term involving the one-step parameter estimates used in the weight matrix can then be obtained as

$$\begin{aligned}\widehat{\text{var}}_c\left(\widehat{\theta}_2\right) = & \frac{1}{N} A^{-1}_{\widehat{\theta}_2,W_N(\widehat{\theta}_1)} C\left(\widehat{\theta}_2\right)' W_N^{-1}\left(\widehat{\theta}_1\right) C\left(\widehat{\theta}_2\right) A^{-1}_{\widehat{\theta}_2,W_N(\widehat{\theta}_1)} \\ & + \frac{1}{N} D_{\widehat{\theta}_2,W_N(\widehat{\theta}_1)} A^{-1}_{\widehat{\theta}_1,W_N} C\left(\widehat{\theta}_1\right)' W_N^{-1} C\left(\widehat{\theta}_2\right) A^{-1}_{\widehat{\theta}_2,W_N(\widehat{\theta}_1)} \\ & + \frac{1}{N} A^{-1}_{\widehat{\theta}_2,W_N(\widehat{\theta}_1)} C\left(\widehat{\theta}_2\right)' W_N^{-1} C\left(\widehat{\theta}_1\right) A^{-1}_{\widehat{\theta}_1,W_N} D'_{\widehat{\theta}_2,W_N(\widehat{\theta}_1)} \\ & + D_{\widehat{\theta}_2,W_N(\widehat{\theta}_1)} \widehat{\text{var}}\left(\widehat{\theta}_1\right) D'_{\widehat{\theta}_2,W_N(\widehat{\theta}_1)} ,\end{aligned}$$

where the jth column of $D_{\widehat{\theta}_2,W_N(\widehat{\theta}_1)}$ is given by

$$D_{\widehat{\theta}_2,W_N(\widehat{\theta}_1)}[.,j] = A^{-1}_{\widehat{\theta}_2,W_N(\widehat{\theta}_1)} C\left(\widehat{\theta}_2\right)' W_N^{-1}\left(\widehat{\theta}_1\right) \frac{\partial W_N(\theta)}{\partial \theta_j}|_{\widehat{\theta}_2} W_N^{-1}\left(\widehat{\theta}_1\right) \overline{g}\left(\widehat{\theta}_2\right) ,$$

and

$$\frac{\partial W_N(\theta)}{\partial \theta_j} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial g_i(\theta)}{\partial \theta_j} g_i(\theta)' + g_i(\theta)\frac{\partial g_i(\theta)'}{\partial \theta_j}\right) .$$

The alternative two-step Wald test that uses a finite sample correction for the asymptotic variance matrix is then defined as

$$\text{Wald}_c = r\left(\widehat{\theta}_2\right)'\left(R'\widehat{\text{var}}_c\left(\widehat{\theta}_2\right)R\right)^{-1} r\left(\widehat{\theta}_2\right) .$$

The term $D_{\theta_0,W(\theta_0)}\left(\widehat{\theta}_1 - \theta_0\right)$ in (18.9) is itself $O_p\left(N^{-1}\right)$ and in this general setting, incorporating non-linear models and/or non-linear moment conditions, whether taking account of it will improve the estimation of the small sample variance substantially depends on the other remainder terms which are of the same order.

18.6.2 Criterion-Based Tests

Using the notation as in Bond and Windmeijer (2005), the standard two-step Wald test can be computed as a criterion difference

$$\text{Wald} = N\left(\overline{g}\left(\widetilde{\theta}_{\widehat{2}}\right)' W_N^{-1}\left(\widehat{\theta}_1\right)\overline{g}\left(\widetilde{\theta}_{\widehat{2}}\right) - \overline{g}\left(\widehat{\theta}_2\right)' W_N^{-1}\left(\widehat{\theta}_1\right)\overline{g}\left(\widehat{\theta}_2\right)\right) ,$$

where $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are the one-step and two-step GMM estimators in the unrestricted model, whereas $\widetilde{\theta}_{\widehat{2}}$ is a two-step GMM estimator in the restricted model, but using as a weight matrix the consistent estimate of Ψ based on the unrestricted one-step GMM estimator, $W_N\left(\widehat{\theta}_1\right)$, see Newey and West (1987).

The LM test can also be computed as a criterion difference

$$\text{LM} = N\left(\overline{g}\left(\widetilde{\theta}_2\right)' W_N^{-1}\left(\widetilde{\theta}_1\right)\overline{g}\left(\widetilde{\theta}_2\right) - \overline{g}\left(\widehat{\widetilde{\theta}}_2\right)' W_N^{-1}\left(\widetilde{\theta}_1\right)\overline{g}\left(\widehat{\widetilde{\theta}}_2\right)\right),$$

where $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ are the one-step and two-step GMM estimators in the restricted model, whereas $\widehat{\widetilde{\theta}}_2$ is a two-step GMM estimator in the unrestricted model, but using as a weight matrix the consistent estimate of Ψ under the null, based on the restricted one-step GMM estimator, $W_N\left(\widetilde{\theta}_1\right)$. The LM test has an asymptotic χ^2_r distribution under the null.

The criterion-based test statistic considered by Bond, Bowsher and Windmeijer (2001) is given by

$$D_{RU} = N\left(\overline{g}\left(\widetilde{\theta}_2\right)' W_N^{-1}\left(\widetilde{\theta}_1\right)\overline{g}\left(\widetilde{\theta}_2\right) - \overline{g}\left(\widehat{\theta}_2\right)' W_N^{-1}\left(\widehat{\theta}_1\right)\overline{g}\left(\widehat{\theta}_2\right)\right).$$

D_{RU} is the "likelihood ratio" test equivalent for GMM, and is the difference between the test statistics for overidentifying restrictions in the restricted and unrestricted models. Under the null, D_{RU} has an asymptotic χ^2_r distribution.

18.6.3 Continuous Updating Estimator

The Continuous Updating Estimator (CUE) is given by

$$\widehat{\theta}_{CU} = \arg\min_{\theta\in\Theta} Q(\theta);$$

$$Q(\theta) = \frac{1}{2}\overline{g}(\theta)' W_N^{-1}(\theta)\overline{g}(\theta),$$

where, as before,

$$W_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} g_i(\theta) g_i(\theta)'$$

and so the CUE minimises the criterion function including the parameters in the weight matrix. The limiting distribution under standard regularity conditions is given by

$$\sqrt{N}\left(\widehat{\theta}_{CU} - \theta_0\right) \to N(0, V);\ V = \left(\Gamma'_{\theta_0}\Psi^{-1}\Gamma_{\theta_0}\right)^{-1}$$

and is the same as the efficient two-step GMM estimator. The asymptotic variance of the CUE is computed as

$$\widehat{\text{var}}\left(\widehat{\theta}_{CU}\right) = \frac{1}{N}\left(C\left(\widehat{\theta}_{CU}\right)' W_N^{-1}\left(\widehat{\theta}_{CU}\right) C\left(\widehat{\theta}_{CU}\right)\right)^{-1},$$

which is used in the calculation of the standard Wald test. Again, it has been shown by e.g. Hansen, Heaton and Yaron (1996) that the asymptotic standard errors are severely downward biased, leading to overrejection of a true null hypothesis using the Wald test.

Newey and Windmeijer (2005) derive the asymptotic distribution of the CUE under many weak instrument asymptotics. In these asymptotics, the number of instruments is allowed to grow with the sample size N, with the increase in number of instruments accompanied by an increase in the concentration parameter. The resulting limiting distribution of the CUE is again the normal distribution, but convergence is at a slower rate than $\sqrt{N}$. The asymptotic variance is in this case larger than the asymptotic variance using conventional asymptotics, and can be estimated consistently by

$$\widehat{\text{var}}\left(\widehat{\theta}_{CU}\right)_c = \frac{1}{N} H^{-1}\left(\widehat{\theta}_{CU}\right) S\left(\widehat{\theta}_{CU}\right)' W_N^{-1}\left(\widehat{\theta}_{CU}\right) S\left(\widehat{\theta}_{CU}\right) H^{-1}\left(\widehat{\theta}_{CU}\right) ,$$

where

$$H(\theta) = \frac{\partial^2 Q(\theta)}{\partial\theta\partial\theta'};\ S(\theta) = (S_1(\theta), S_2(\theta), \ldots, S_k(\theta))$$

$$S_j(\theta) = \left(\frac{\partial \overline{g}(\theta)}{\partial \theta_j} - \Lambda_j(\theta) W_N^{-1}(\theta)\overline{g}(\theta)\right)$$

$$\Lambda_j(\theta) = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial g_i(\theta)}{\partial \theta_j} g_i(\theta)' .$$

Here, unlike the usual asymptotics, the middle matrix $S\left(\widehat{\theta}_{CU}\right)' W_N^{-1}\left(\widehat{\theta}_{CU}\right) S\left(\widehat{\theta}_{CU}\right)$ estimates a different, larger object than the Hessian. Also, the use of the Hessian is important, as the more common formula $C\left(\widehat{\theta}_{CU}\right)' W_N^{-1}\left(\widehat{\theta}_{CU}\right) C\left(\widehat{\theta}_{CU}\right)$ has extra random terms that are eliminated in the Hessian under the alternative asymptotics.

Hansen, Heaton and Yaron (1996) proposed the use of a criterion-based test similar to D_{RU}, but based on the CUE. Their test statistic D_{RU}^{CU} is defined as

$$D_{RU}^{CU} = N\left(Q\left(\widetilde{\theta}_{CU}\right) - Q\left(\widehat{\theta}_{CU}\right)\right) ,$$

where $\widehat{\theta}_{CU}$ and $\widetilde{\theta}_{CU}$ are the CUEs for the unrestricted and restricted models respectively. Under the null, D_{RU}^{CU} has an asymptotic χ_r^2 distribution.

18.6.4 Monte Carlo Results

In this section we will illustrate the finite sample performance of the GMM estimators and the test statistics as discussed in the previous sections by means of a small Monte Carlo study. The data generating process is given by

$$
\begin{aligned}
&y_{it} \sim \text{Poisson}\left(\exp\left(x_{it}\beta + \eta_i + \varepsilon_{it}\right)\right) \\
&x_{it} = \rho x_{it-1} + \delta\eta_i + \theta\varepsilon_{it-1} + \omega_{it} \\
&\eta_i \sim N\left(0, \sigma_\eta^2\right);\ \varepsilon_{it} \sim N\left(0, \sigma_\varepsilon^2\right);\ \omega_{it} \sim N\left(0, \sigma_\omega^2\right), \\
&\beta = 0.5;\ \delta = 0.1;\ \theta = 0.3;\ \sigma_\eta^2 = 0.3;\ \sigma_\varepsilon^2 = 0.3;\ \sigma_\omega^2 = 0.25 \\
&\rho = \{0.5, 0.8\}\ .
\end{aligned}
$$

The dependent variable is a count variable, generated from the Poisson distribution with unobserved fixed normally distributed heterogeneity η_i and further idiosyncratic normally distributed heterogeneity ε_{it}. The x_{it} are correlated with the η_i and ε_{it-1} and are therefore predetermined.

Table 18.1 presents estimation results from 10,000 Monte Carlo replications for the one- and two-step GMM estimators as well as the continuous updating estimator for $T = 6$, $N = 250$ and $\rho = 0.5$ or $\rho = 0.8$, using the moment conditions (18.6) as proposed by Chamberlain (1992). The instruments set is given by

$$
Z_i = \begin{bmatrix} x_{i1} & & & & \\ & x_{i1} & x_{i2} & & \\ & & & \ddots & \\ & & & & x_{i1} \ \cdots \ x_{iT-1} \end{bmatrix}
$$

and hence there are a total of 15 moment conditions.The one-step GMM estimator uses $W_N = \frac{1}{N}\sum_{i=1}^N Z_i' Z_i$ as the weight matrix.

When $\rho = 0.5$, the instruments are quite strong. The one-step GMM estimator, denoted GMM1 in the table, has a moderate downward bias of -0.0408. Its standard deviation is 0.1053, which seems well approximated by the asymptotic standard error. The mean of the estimated standard errors is equal to 0.1031. The two-step GMM estimator, denoted GMM2, has a smaller bias of -0.0211 and a smaller standard deviation of 0.0803, representing a substantial efficiency gain with more than a 23% reduction in standard deviation. In contrast to the one-step esti-

Table 18.1 Estimation results

$\rho = 0.5$	Mean Bias	St Dev	Se	Sec	Median Bias	IQR
GMM1	−0.0408	0.1053	0.1031		−0.0409	0.1381
GMM2	−0.0211	0.0803	0.0652	0.0799	−0.0209	0.1077
CUE	0.0043	0.0904	0.0652	0.0918	0.0024	0.1165
$\rho = 0.8$						
GMM1	−0.1136	0.2094	0.1773		−0.0974	0.2435
GMM2	−0.0537	0.1335	0.0908	0.1365	−0.0498	0.1558
CUE	0.0033	0.1885	0.0879	0.1459	0.0029	0.1742

Note: $T = 6$, $N = 250$, $\beta = 0.5$, 10,000 replications, sec denotes finite sample corrected standard errors for GMM2 and those resulting from many weak instrument asymptotics for CUE, IQR= Inter Quartile Range

mator, the mean of the usual asymptotic standard errors is 19% smaller than the standard deviation. However, taking account of the extra variation due to the presence of the one-step estimates in the weight matrix results in finite sample corrected standard errors with a mean of 0.0799, which is virtually identical to the standard deviation. The CUE has a very small bias of 0.0043, with a standard deviation of 0.0904, which is larger than that of the two-step GMM estimator, but smaller than that of the one-step estimator. The mean of the usual asymptotic standard errors is exactly the same as that of the two-step GMM estimator and in this case it is almost 28% smaller than the standard deviation. The standard errors resulting from the many weak instruments asymptotics have a mean of 0.0918, which is virtually the same as the standard deviation.

Figure 18.1 shows p-value plots for the hypothesis $H_0 : \rho = 0.5$, comparing nominal size with rejection frequencies. The various Wald tests are denoted W_1, W_2, W_{2C}, W_{CU} and W_{CUC} based on one-step GMM, two-step GMM with usual standard errors, two-step GMM with finite sample corrected standard errors, CUE and CUE with standard errors resulting from the many weak instruments asymptotics, respectively. As expected, W_2 and W_{CU} overreject the null hypothesis substantially. W_1 and W_{2C} perform much better, but are still moderately oversized due to the bias of the estimators. W_{CUC} has a very good performance in terms of size of the test, the rejection frequencies being very close to the 45^o line. The two-step GMM based LM and D_{RU} tests also perform very well, their p-value plots being quite similar to

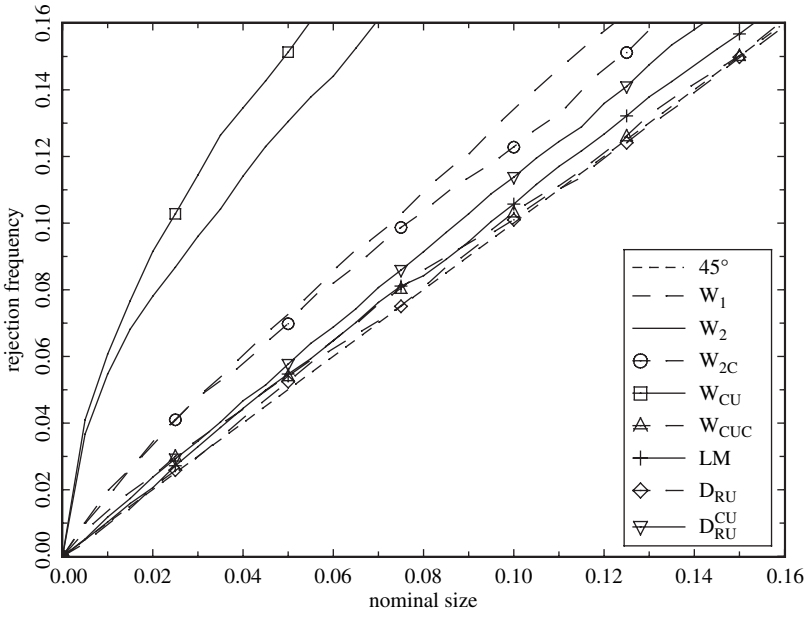

Fig. 18.1 *P*-value plot, $H_0 : \rho = 0.5$

that of W_{CUC}. Finally, the CUE based D_{RU}^{CU} performs well, but tends to overreject moderately at higher values of the nominal size.

When $\rho = 0.8$, the instruments become weaker as the x_{it} series become more persistent. The one-step GMM estimator now has a quite large downward bias of -0.1136. Its standard deviation is 0.2094, which is now less well approximated by the asymptotic standard error, with the mean of the estimated standard errors being equal to 0.1773. The two-step GMM estimator has a smaller, but still quite substantial bias of -0.0537 and a smaller standard deviation of 0.1335. The mean of the usual asymptotic standard errors is 0.0908, again substantially smaller than the standard deviation. The mean of the finite sample corrected standard errors is 0.1365, which is again very close to the standard deviation. The CUE, also with these weaker instruments, has a very small bias of 0.0033, with a standard deviation of 0.1885. In this case the so-called no moment-problem starts to become an issue for the CUE, though, with some outlying estimates inflating the standard deviation, see Guggenberger (2005). It is therefore better to look at the median bias and inter quartile range (IQR) in this case, which shows that the CUE is median unbiased with an IQR which is only slightly larger than that of the two-step GMM estimator, 0.1742 versus 0.1558 respectively.

Figure 18.2 shows p-value plots for the hypothesis $H_0 : \rho = 0.8$. W_2 and W_{CU} overreject the null hypothesis even more than when $\rho = 0.5$. W_1 performs better, but is still substantially oversized. W_{2C} and W_{CUC} perform quite well and quite similar, slightly overrejecting the null. The two-step GMM based LM and D_{RU} are again the

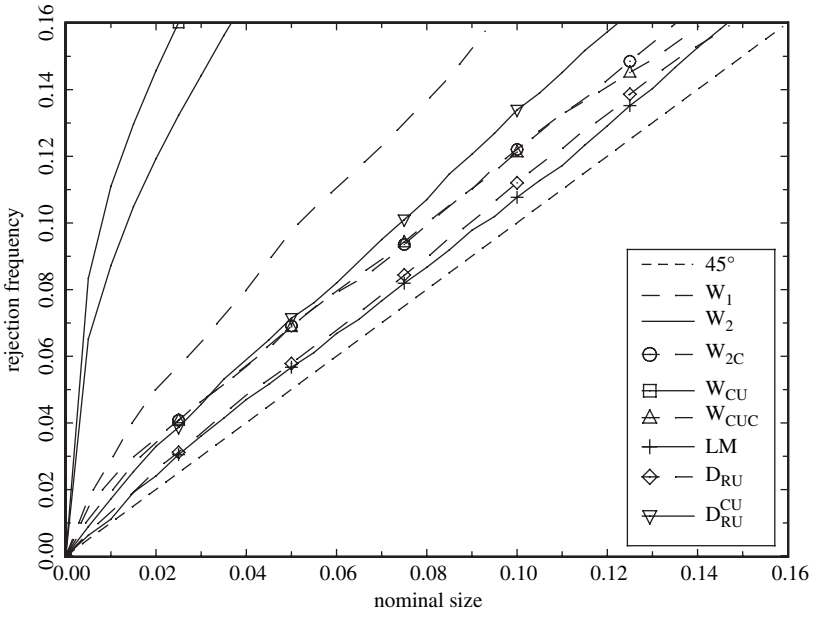

Fig. 18.2 P-value plot, $H_0 : \rho = 0.8$

best performers in terms of size, whereas the CUE based D_{RU}^{CU} performs worse than W_{2C} and W_{CUC}.

Summarising, it is clear that use of the finite sample corrected standard errors for the two-step GMM estimator and the standard errors from the many weak instrument asymptotics for the CUE improve the size performance of the Wald tests for these estimators considerably. The simple criterion based D_{RU} test performs very well in these examples, as was the case in Bond and Windmeijer (2005) for linear panel data models.

References

Angrist J.D., 2001, Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice, Journal of Business & Economic Statistics 19, 2–16.

Aptech Systems, Inc., 2005, GAUSS, Advanced Mathematical and Statistical System, Black Diamond, WA, USA.

Blundell R. and S.R. Bond, 1998, Initial conditions and moment restrictions in dynamic panel data models, Journal of Econometrics 87, 115–144.

Blundell R., R. Griffith and J. van Reenen, 1999, Market share, market value and innovation in a panel of British manufacturing firms, Review of Economic Studies 66, 529–554.

Blundell R., R. Griffith and F. Windmeijer, 2002, Individual effects and dynamics in count data models, Journal of Econometrics 108, 113–131.

Bond S.R., C. Bowsher and F. Windmeijer, 2001, Criterion-Based inference for GMM in autoregressive panel data models, Economics Letters 73, 379–388.

Bond S.R. and F. Windmeijer, 2005, Reliable inference for GMM estimators? Finite sample procedures in linear panel data models, Econometric Reviews 24, 1–37.

Chamberlain G., 1984, Panel data. In: Griliches, Z. and M. Intriligator (Eds.), Handbook of Econometrics, North Holland, Amsterdam.

Chamberlain G., 1992, Comment: Sequential moment restrictions in panel data, Journal of Business & Economic Statistics 10, 20–26.

Cincera M., 1997, Patents, R&D, and technological spillovers at the firm level: some evidence from econometric count models for panel data, Journal of Applied Econometrics 12, 265–280.

Crépon B. and E. Duguet, 1997, Estimating the innovation function from patent numbers: GMM on count panel data, Journal of Applied Econometrics 12, 243–263.

Greene W.H., 2005, LIMDEP 8.0, Econometric Software, Inc., Plainview, NY, USA.

Guggenberger P., 2005, Monte-carlo evidence suggesting a no moment problem of the continuous updating estimator, Economics Bulletin 3, 1–6.

Hall B. and C. Cummins, 2005, TSP 5.0, TSP International, Palo Alto, CA, USA.

Hansen L.P., 1982, Large sample properties of generalized method of moments estimators, Econometrica 50, 1029–1054.

Hansen L.P., J. Heaton and A. Yaron, 1996, Finite-sample properties of some alternative GMM estimators, Journal of Business & Economic Statistics 14, 262–280.

Hausman J., B. Hall and Z. Griliches, 1984, Econometric models for count data and an application to the patents-R&D relationship, Econometrica 52, 909–938.

Kelly M., 2000, Inequality and crime, The Review of Economics and Statistics 82, 530–539.

Kitazawa Y., 2000, TSP procedures for count panel data estimation, Kyushu Sangyo University.

Lancaster T., 2002, Orthogonal parameters and panel data, Review of Economic Studies 69, 647–666.

Kim J. and G. Marschke, 2005, Labor mobility of scientists, technological diffusion and the firm's patenting decision, The RAND Journal of Economics 36, 298–317.
Manning W.G., A. Basu and J. Mullahy, 2005, Generalized modeling approaches to risk adjustment of skewed outcomes data, Journal of Health Economics 24, 465–488.
Montalvo J.G., 1997, GMM estimation of count-panel-data models with fixed effects and predetermined instruments, Journal of Business and Economic Statistics 15, 82–89.
Mullahy J., 1997, Instrumental variable estimation of Poisson regression models: application to models of cigarette smoking behavior, Review of Economics and Statistics 79, 586–593.
Newey W.K. and R.J. Smith, 2004, Higher order properties of GMM and generalized empirical likelihood estimators, Econometrica 72, 219–255.
Newey W.K. and K.D. West, 1987, Hypothesis testing with efficient method of moments estimation, International Economic Review 28, 777–787.
Newey W.K. and F. Windmeijer, 2005, GMM with many weak moment conditions, cemmap Working Paper No. CWP18/05.
Romeu A., 2004, *ExpEnd*: Gauss code for panel count data models, Journal of Applied Econometrics 19, 429–434.
Salomon R.M. and J.M. Shaver, 2005, Learning by exporting: new insights from examining firm innovation, Journal of Economics and Management Strategy 14, 431–460.
Santos Silva J.M.C. and S. Tenreyro, 2006, The log of gravity, The Review of Economics and Statistics, 88, 641–658.
Schellhorn M., 2001, The effect of variable health insurance deductibles on the demand for physician visits, Health Economics 10, 441– 456.
Vera-Hernandez, A.M., 1999, Duplicate coverage and demand for health care. The case of Catalonia, Health Economics 8, 579–598.
Windmeijer F., 2000, Moment conditions for fixed effects count data models with endogenous regressors, Economics Letters 68, 21–24.
Windmeijer F., 2002, *ExpEnd*, a Gauss programme for non-linear GMM estimation of exponential models with endogenous regressors for cross section and panel data, c*e*mmap Working Paper No. CWP14/02.
Windmeijer F., 2005, A finite sample correction for the variance of linear efficient two-step GMM estimators, Journal of Econometrics 126, 25–517.
Windmeijer F. and J.M.C. Santos Silva, 1997, Endogeneity in count data models: an application to demand for health care, Journal of Applied Econometrics 12, 281–294.
Wooldridge J.M., 1991, Multiplicative panel data models without the strict exogeneity assumption, MIT Working Paper No. 574.
Wooldridge J.M., 1997, Multiplicative panel data models without the strict exogeneity assumption, Econometric Theory 13, 667–678.
Wooldridge J.M., 1999, Distribution-free estimation of some nonlinear panel data models, Journal of Econometrics 90, 77–97.

Chapter 19
Spatial Panel Econometrics

Luc Anselin, Julie Le Gallo and Hubert Jayet

19.1 Introduction

Spatial econometrics is a subfield of econometrics that deals with the incorporation of *spatial effects* in econometric methods (Anselin, 1988a). Spatial effects may result from *spatial dependence*, a special case of cross-sectional dependence, or from *spatial heterogeneity*, a special case of cross-sectional heterogeneity. The distinction is that the *structure* of the dependence can somehow be related to location and distance, both in a geographic space as well as in a more general economic or social network space. Originally, most of the work in spatial econometrics was inspired by research questions arising in regional science and economic geography (early reviews can be found in, among others, Paelinck and Klaassen, 1979; Cliff and Ord, 1981; Upton and Fingleton, 1985; Anselin, 1988a; Haining, 1990; Anselin and Florax, 1995). However, more recently, spatial (and social) interaction has increasingly received more attention in mainstream econometrics as well, both from a theoretical as well as from an applied perspective (see the recent reviews and extensive references in Anselin and Bera, 1998; Anselin, 2001b, 2002; Florax and Van Der Vlist, 2003; and Anselin et al., 2004).

The central focus in spatial econometrics to date has been the single equation cross-sectional setting. However, as Arrelano argues in the introduction to his recent panel data econometrics text, "the field [of econometrics of panel data] has expanded to cover almost any aspect of econometrics" (Arellano, 2003, p. 2). It is therefore not surprising that this has included spatial econometrics as well. For

Luc Anselin
School of Geographical Sciences, Arizona State University, Tempe, AZ 85287, USA,
e-mail: luc.anselin@asu.edu

Julie Le Gallo
CRESE, Université de Franche-Comté, 45D Avenue de l'Observatoire, 25030 Besançon Cedex, France, e-mail: jlegallo@univ-fcomte.fr

Hubert Jayet
EQUIPPE, University of Science and Technology of Lille, Faculty of Economics and Social Sciences, 59655 Villeneuve d'Ascq Cedex, France, e-mail: Hubert.Jayet@univ-lille1.fr

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

example, the second edition of Baltagi's well known panel data text now includes a brief discussion of *spatial panels* (Baltagi, 2001, pp. 195–197), and an increasing number of papers are devoted to the topic (see the reviews in Anselin, 2001b; Elhorst, 2001, 2003, as well as the recent papers by Baltagi et al., 2007, 2006; Kapoor et al., 2007; and Pesaran, 2004; among others).

In this chapter, we review and organize this recent literature and emphasize a range of issues pertaining to the specification, estimation and diagnostic testing for spatial effects in panel data models. Since this encompasses a large and rapidly growing literature, we limit our attention to models with continuous dependent variables,[1] and to a design where the cross-sectional dimension (N) vastly exceeds the time dimension ($N \gg T$). We also avoid duplication by excluding aspects of the standard treatment of heterogeneity and dependence in panel data models, as well as the case where cross-sectional dependence is modeled by relying on the time dimension (e.g., as in the classic SURE case with fixed N, and some more recent extensions, such as Chen and Conley, 2001).

The chapter is organized into five remaining sections. First, we define the notion of *spatial effects* more precisely and provide a brief outline of how the traditional cross-sectional models can be extended to panel data model specifications. Next, we consider this more closely and develop a taxonomy of space-time models. We then turn to the issues of model estimation and diagnostic testing. We close with some concluding remarks.

19.2 Spatial Effects

As a point of departure, consider a simple pooled linear regression model:

$$y_{it} = x_{it}\beta + \varepsilon_{it}, \tag{19.1}$$

where i is an index for the cross-sectional dimension, with $i = 1, \ldots, N$, and t is an index for the time dimension, with $t = 1, \ldots, T$.[2] Using customary notation, y_{it} is an observation on the dependent variable at i and t, x_{it} a $1 \times K$ vector of observations on the (exogenous) explanatory variables, β a matching $K \times 1$ vector of regression coefficients, and ε_{it} an error term.

Given our interest in spatial effects, the observations will be stacked as successive cross-sections for $t = 1, \ldots, T$, referred to as y_t (a $N \times 1$ vector of cross-sectional observations for time period t), X_t (a $N \times K$ matrix of observations on a cross-section of the explanatory variables for time period t) and ε_t (a $N \times 1$ vector of cross-sectional

[1] The treatment of spatial effects in panel data models with discrete dependent variables is still in its infancy.

[2] Note that we couch the discussion using "time" as the second dimension for the sake of simplicity. In general, it is also possible to have the second dimension reflect another cross-sectional characteristic, such as an industry sector, and. along the same lines, extension to higher order panel structures are possible as well.

disturbances for time period t). In stacked form, the simple pooled regression then becomes:

$$y = X\beta + \varepsilon, \tag{19.2}$$

with y as a $NT \times 1$ vector, X as a $NT \times K$ matrix and ε as a $NT \times 1$ vector.

In general, spatial dependence is present whenever correlation across cross-sectional units is non-zero, and the pattern of non-zero correlations follows a certain spatial *ordering*. When little is known about the appropriate spatial ordering, spatial dependence is reduced to simple cross-sectional dependence. For example, the error terms are spatially correlated when $\mathrm{E}[\varepsilon_{it}\varepsilon_{jt}] \neq 0$, for a given t and $i \neq j$, and the non-zero covariances conform to a specified *neighbor* relation. Note how the correlation is purely cross-sectional in that it pertains to the same time period t.

The neighbor relation is expressed by means of a so-called spatial weights matrix. We will briefly review the concept of spatial weights (and the associated spatial lag operator) and outline two classes of specifications for models with spatial dependence. In one, the spatial correlation pertains to the dependent variable, in a so-called *spatial lag* model, in the other it affects the error term, a so-called *spatial error* model. The two specifications can also be combined, resulting in so-called higher order spatial models. While these models and terms are by now fairly familiar in the spatial econometric literature, we thought it useful to briefly review them and to illustrate how they may be incorporated into a panel data setting.[3]

The second class of spatial effects, spatial heterogeneity, is a special case of the observed and unobserved heterogeneity which is treated prominently in the mainstream panel data econometrics literature. For example, a heterogeneous panel would relax the constant regression coefficient in (19.1), and replace it by:

$$y_{it} = x_{it}\beta_i + \varepsilon_{it} \ ,$$

where the β_i is a $K \times 1$ vector of regression coefficients specific to the cross-sectional unit i.

This heterogeneity becomes *spatial* when there is a structure to the variability across the i that is driven by spatial variables, such as location, distance or region. In the spatial literature, discrete spatial variability is referred to as *spatial regimes* (Anselin, 1988a). The continuous case can be modeled as a special form of random coefficient variation (where the covariance shows a spatial pattern), or deterministically, as a function of extraneous variables (so-called spatial expansion, e.g., Casetti, 1997), or as a special case of local regression models (so-called geographically weighted regression, Fotheringham et al., 2002). Neither of these has seen application in panel data contexts.[4]

---

[3] A more extensive technical review can be found in Anselin and Bera (1998).

[4] In the literature of spatial statistics, spatially varying coefficients are treated in (Bayesian) hierarchical models (Gelfand et al., 2003; Gamerman et al., 2003).

Since most econometric aspects of spatial heterogeneity can be handled by means of the standard panel data methods, we will focus the discussion that follows on spatial dependence and will only consider the heterogeneity when it is relevant to the modeling of the dependence.

19.2.1 Spatial Weights and Spatial Lag Operator

A spatial weights matrix W is a $N \times N$ positive matrix in which the rows and columns correspond to the cross-sectional observations. An element w_{ij} of the matrix expresses the prior strength of the interaction between location i (in the row of the matrix) and location j (column). This can be interpreted as the presence and strength of a link between nodes (the observations) in a network representation that matches the spatial weights structure. In the simplest case, the weights matrix is binary, with $w_{ij} = 1$ when i and j are neighbors, and $w_{ij} = 0$ when they are not. By convention, the diagonal elements $w_{ii} = 0$. For computational simplicity and to aid in the interpretation of the spatial variables, the weights are almost always standardized such that the elements in each row sum to 1, or, $w^s_{ij} = w_{ij}/\sum_j w_{ij}$.[5] A side effect of this standardization is that the sum of all elements in W equals N, the number of cross-sectional observations. Whereas the original weights are often symmetric, the row-standardized form is no longer, which is an unusual complication with significant computational consequences.

The specification of the spatial weights is an important problem in applied spatial econometrics.[6] Unless the weights are based on a formal theoretical model for social or spatial interaction, their specification is often ad hoc. In practice, the choice is typically driven by geographic criteria, such as contiguity (sharing a common border) or distance, including nearest neighbor distance (for examples and further discussion, see, e.g., Cliff and Ord, 1981, pp. 17–19; Anselin, 1988a, Chap. 3).

Generalizations that incorporate notions of "economic" distance are increasingly used as well (e.g., Case et al., 1993; Conley and Ligon, 2002; Conley and Topa, 2002). A slightly different type of economic weights are so-called block weights, where all observations in the same region are considered to be neighbors (and not only the adjoining observations). More formally, if there are N_g units in a block (such as counties in a state), they are all considered to be neighbors, and the spatial weights equal $1/(N_g - 1)$ for all observations belonging to the same block (see, e.g., Case, 1991, 1992; and, more recently, Lee, 2002).

So far, the weights considered were purely cross-sectional. To extend their use in a panel data setting, they are assumed to remain constant over time.[7]

[5] In what follows, we will use the symbol W for the spatial weights and assume row-standardization.

[6] An extensive discussion of spatial weights is outside the scope of this chapter. For a detailed assessment of technical issues, see the recent review papers by Anselin and Bera (1998), and Anselin (2002).

[7] Since the spatial weights enter into a model premultiplied by a scalar parameter, changes in the interaction structure over time can be accounted for by allowing this parameter to vary. Alternatively,

Using the subscript to designate the matrix dimension, with W_N as the weights for the cross-sectional dimension, and the observations stacked as in (19.2), the full $NT \times NT$ weights matrix then becomes:

$$W_{NT} = I_T \otimes W_N, \tag{19.3}$$

with I_T as an identity matrix of dimension T.

Unlike the time series case, where "neighboring" observations are directly incorporated into a model specification through a shift operator (e.g., $t-1$), this is not unambiguous in a two dimensional spatial setting. For example, observations for irregular spatial units, such as counties or census tracts, typically do not have the same number of neighbors, so that a spatial shift operator cannot be implemented. Instead, in spatial econometrics, the neighboring observations are included through a so-called *spatial lag* operator, more akin to a distributed lag than a shift (Anselin, 1988a). In essence, a spatial lag operator constructs a new variable that consists of the weighted average of the neighboring observations, with the weights as specified in W. More formally, for a cross-sectional observation i for variable z, the spatial lag would be $\sum_j w_{ij} z_j$. In most applications, the bulk of the row elements in w_{ij} are zero (resulting in a *sparse* structure for W) so that in effect the summation over j only incorporates the "neighbors," i.e., those observations for which $w_{ij} \neq 0$. In matrix notation, this corresponds to the matrix operation $W_N y_t$, in which the $N \times N$ cross-sectional weights matrix is post-multiplied by a $N \times 1$ vector of cross-sectional observations for each time period $t = 1, \ldots, T$.

Spatial variables are included into a model specification by applying a spatial lag operator to the dependent variable, to the explanatory variables, or to the error term. A wide range of models for local and global spatial externalities can be specified in this manner (for a review, see Anselin, 2003). This extends in a straightforward manner to the panel data setting, by applying the $NT \times NT$ weights from (19.3) to the stacked y, X or ε from (19.2).

More precisely, in the same notation as above, a vector of spatially lagged dependent variables follows as:

$$Wy = W_{NT} y = (I_T \otimes W_N) y, \tag{19.4}$$

a matrix of spatially lagged explanatory variables as:

$$WX = W_{NT} X = (I_T \otimes W_N) X \ ,$$

and a vector of spatially lagged error terms as:

$$W\varepsilon = W_{NT} \varepsilon = (I_T \otimes W_N) \varepsilon \ .$$

---

but less tractable, would be to let the weights vary and keep the parameter constant. Obviously, letting both parameter and weights vary over time would lead to problems with identification and interpretation (for example, see Druska and Horrace, 2004).

The incorporation of these spatial lags into a regression specification is considered next.

19.2.2 Spatial Lag Model

A spatial lag model, or, mixed regressive spatial autoregressive model, includes a spatially lagged dependent variable on the RHS of the regression specification (Anselin, 1988a). While usually applied in a pure cross-sectional setting, it can easily be extended to panel models. Using the stacked (19.2) and the expression for the spatial lag from (19.4), this yields:

$$y = \rho(I_T \otimes W_N)y + X\beta + \varepsilon \,, \tag{19.5}$$

where ρ is the spatial autoregressive parameter, and the other notation is as before.

In a cross-section, a spatial lag model is typically considered as the formal specification for the equilibrium outcome of a spatial or social interaction process, in which the value of the dependent variable for one agent is jointly determined with that of the neighboring agents.[8] This model is increasingly applied in the recent literature on social/spatial interaction, and is used to obtain empirical estimates for the parameters of a spatial reaction function (Brueckner, 2003) or social multiplier (Glaeser et al., 2002). It should be noted that other formulations to take into account social interaction have been suggested as well (e.g., Manski, 2000; Brock and Durlauf, 2001) mostly in the context of discrete choice. The modeling of complex neighborhood and network effects (e.g., Topa, 2001) requires considerable attention to identification issues, maybe best known from the work of Manski on the "reflection problem" (Manski, 1993). Because of this theoretical foundation, the choice of the weights in a spatial lag model is very important.

At first sight, the extension of the spatial lag model to a panel data context would presume that the equilibrium process at hand is stable over time (constant ρ and constant W). However, the inclusion of the time dimension allows much more flexible specifications, as outlined in Sect. 19.3.

The essential econometric problem in the estimation of (19.5) is that, unlike the time series case, the spatial lag term is *endogenous*. This is the result of the two-directionality of the neighbor relation in space ("I am my neighbor's neighbor") in contrast to the one-directionality in time dependence (for details, see Anselin and Bera, 1998). The consequence is a so-called *spatial multiplier* (Anselin, 2003) which formally specifies how the joint determination of the values of the dependent

[8] In spatial statistics, the preferred perspective is that of a conditional process, which is geared to spatial prediction (see Cressie, 1993; Stein, 1999; and for a discussion of the implications in a spatial econometric context, Anselin, 2002). Rather than specifying the joint distribution of all the y_i in the system, each y_i is modeled conditional upon the y_j for the neighbors. For detailed discussion, see the previous references.

variables in the spatial system is a function of the explanatory variables and error terms at all locations in the system.

The extent of the joint determination of values in the system can be seen by expressing (19.5) as a reduced form:

$$y = \left[I_T \otimes (I_N - \rho W_N)^{-1}\right] X\beta + \left[I_T \otimes (I_N - \rho W_N)^{-1}\right] \varepsilon \,, \tag{19.6}$$

with the subscripts indicating the dimensions of the matrices. The inverse matrix expression can be expanded and considered one cross-section at a time, due to the block-diagonal structure of the inverse. A a result, for each $N \times 1$ cross-section at time $t = 1, \ldots, T$:

$$y_t = X_t\beta + \rho W_N X_t\beta + \rho^2 W_N^2 X_t\beta + \ldots + \varepsilon_t + \rho W_N \varepsilon_t + \rho^2 W_N^2 \varepsilon_t \ldots$$

The implication of this reduced form is that the spatial distribution of the y_{it} in each cross-section is determined not only by the explanatory variables and associated regression coefficients at each location ($X_t\beta$), but also by those at neighboring locations, albeit subject to a distance decay effect (the increasing powers of ρ and W_N). In addition, the unobserved factors contained in the error term are not only relevant for the location itself, but also for the neighboring locations ($W_N\varepsilon$), again, subject to a distance decay effect. Note that in the simple pooled model, this spatial multiplier effect is contained within each cross-section and does not spill over into other time periods.[9]

The presence of the spatially lagged errors in the reduced form illustrates the joint dependence of the $W_N y_t$ and ε_t in each cross-section. In model estimation, this simultaneity must be accounted for through instrumentation (IV and GMM estimation) or by specifying a complete distributional model (maximum likelihood estimation).

Even without a solid theoretical foundation as a model for social/spatial interaction, a spatial lag specification may be warranted to spatially detrend the data. This is referred to as a *spatial filter*:

$$[I_T \otimes (I_N - \rho W_N)]\, y = X\beta + \varepsilon, \tag{19.7}$$

with the LHS as a new dependent variable from which the effect of spatial autocorrelation has been eliminated. In contrast to time series, a simple detrending using $\rho = 1$ is not possible, since that value of ρ is not in the allowable parameter space.[10] As a consequence, the parameter ρ must be estimated in order for the spatial filtering to be operational (see Anselin, 2002).

[9] This can be relaxed in more flexible space-time models, see, for example, Sect. 19.3.3.1.

[10] For row-standardized weights, $\rho = 1$ violates a standard regularity condition for spatial models that requires that the inverse $(I - \rho W_N)^{-1}$ exists (Kelejian and Prucha, 1999).

19.2.3 Spatial Error Model

In contrast to the spatial lag model, a spatial error specification does not require a theoretical model for spatial/social interaction, but, instead, is a special case of a non-spherical error covariance matrix. An unconstrained error covariance matrix at time t, $\mathrm{E}[\varepsilon_{it}\varepsilon_{jt}], \forall\, i \neq j$ contains $N \times (N-1)/2$ parameters. These are only estimable for small N and large T, and provided they remain constant over the time dimension. In the panel data setting considered here, with $N \gg T$, structure must be imposed in order to turn the covariance matrix into a function of a manageable set of parameters.

Four main approaches have been suggested to provide the basis for a parsimonious covariance structure: direct representation, spatial error processes, spatial error components, and common factor models. Each will be reviewed briefly.

19.2.3.1 Direct Representation

The direct representation approach has its roots in the geostatistical literature and the use of theoretical variogram and covariogram models (Cressie, 1993). It consists of specifying the covariance between two observations as a *direct* function of the distance that separates them, $\forall\, i \neq j$ and $t = 1, \ldots, T$:

$$\mathrm{E}[\varepsilon_{it}\varepsilon_{jt}] = \sigma^2 f(\tau, d_{ij}), \tag{19.8}$$

where τ is a parameter vector, d_{ij} is the (possibly economic) distance between observation pairs i, j, σ^2 is a scalar variance term, and f is a suitable distance decay function, such as a negative exponential.[11] The parameter space for τ should be such that the combination of functional form and the distance metric ensures that the resulting covariance matrix is positive definite (for further discussion, see, e.g, Dubin, 1988).

An extension to a panel data setting is straightforward. With $\sigma^2\Omega_{t,N}$ as the error covariance matrix that results from applying the function (19.8) to the $N \times 1$ cross-sectional error vector in time period t, the overall $NT \times NT$ error variance-covariance matrix Σ_{NT} becomes a block diagonal matrix with the $N \times N$ variance matrix for each cross-section on the diagonal.[12] However, as specified, the function (19.8) does not vary over time, so that the result can be expressed concisely as:

$$\Sigma_{NT} = \sigma^2 \left[I_T \otimes \Omega_N \right] ,$$

with $\Omega_{t,N} = \Omega_N\ \forall t$.[13]

[11] For the sake of simplicity, we use a homoskedastic model with constant variance across all time periods. This restriction can be readily relaxed. Similarly, the assumption of isotropy (only distance matters, not direction) may be relaxed by including separate functions to account for directional effects.

[12] In the notation that follows, we use the subscripts T, N and NT to refer to the dimension of the matrix, and the subscript t to refer to the cross-section at time t.

[13] Note that this simplification only holds in the strictly homogeneous case with $\sigma_t^2 = \sigma^2\ \forall t$.

19.2.3.2 Spatial Error Processes

Whereas the direct representation approach requires a distance metric and functional form for the distance decay between a pair of observations, spatial error processes are based on a formal relation between a location and its *neighbors*, using a spatial weights matrix. The error covariance structure can then be derived for each specified process, but typically the *range* of neighbors specified in the model is different from the range of spatial dependence in the covariance matrix. This important aspect is sometimes overlooked in empirical applications.

In analogy to time series analysis, the two most commonly used models for spatial processes are the autoregressive and the moving average (for extensive technical discussion, see Anselin, 1988a; Anselin and Bera, 1998; Anselin, 2003, and the references cited therein).

A *spatial autoregressive* (SAR) specification for the $N \times 1$ error vector ε_t in period $t = 1, \ldots, T$, can be expressed as:

$$\varepsilon_t = \theta W_N \varepsilon_t + u_t \ ,$$

where W_N is a $N \times N$ spatial weights matrix (with the subscript indicating the dimension), θ is the spatial autoregressive parameter, and u_t is a $N \times 1$ idiosyncratic error vector, assumed to be distributed independently across the cross-sectional dimension, with constant variance σ_u^2.

Continuing in matrix notation for the cross-section at time t, it follows that:

$$\varepsilon_t = (I_N - \theta W_N)^{-1} u_t \ ,$$

and hence the error covariance matrix for the cross-section at time t becomes:

$$\Omega_{t,N} = \mathrm{E}[\varepsilon_t \varepsilon_t'] = \sigma_u^2 (I_N - \theta W_N)^{-1} \left(I_N - \theta W_N'\right)^{-1} \ ,$$

or, in a simpler notation, with $B_N = I_N - \theta W_N$:

$$\Omega_{t,N} = \sigma_u^2 (B_N' B_N)^{-1} \ .$$

As before, in this homogeneous case, the cross-sectional covariance does not vary over time, so that the full $NT \times NT$ covariance matrix follows as:

$$\Sigma_{NT} = \sigma_u^2 \left[I_T \otimes (B_N' B_N)^{-1}\right] . \tag{19.9}$$

Note that for a row-standardized weights matrix, B_N will not be symmetric. Also, even though W_N may be sparse, the inverse term $(B_N' B_N)^{-1}$ will not be sparse and suggests a much wider range of spatial *covariance* than specified by the non-zero elements of the weights matrix. In other words, the spatial covariance structure induced by the SAR model is *global*.

A *spatial moving average* (SMA) specification for the $N \times 1$ error vector ε_t in period $t = 1, \ldots, T$, can be expressed as:

$$\varepsilon_t = \gamma W_N u_t + u_t \ ,$$

where γ is the moving average parameter, and the other notation is as before. In contrast to the SAR model, the variance covariance matrix for an error SMA process does not involve a matrix inverse:

$$\Omega_{t,N} = \mathrm{E}[\varepsilon_t \varepsilon_t'] = \sigma_u^2 \left[I_N + \gamma(W_N + W_N') + \gamma^2 W_N W_N' \right], \tag{19.10}$$

and, in the homogenous case, the overall error covariance matrix follows directly as:

$$\Sigma_{NT} = \sigma_u^2 \left(I_T \otimes \left[I_N + \gamma(W_N + W_N') + \gamma^2 W_N W_N' \right] \right) \ .$$

Aso, in contrast to the SAR model, the spatial covariance induced by the SMA model is *local*.[14]

19.2.3.3 Spatial Error Components

A spatial error components specification (SEC) was suggested by Kelejian and Robinson as an alternative to the SAR and SMA models (Kelejian and Robinson, 1995; Anselin and Moreno, 2003). In the SEC model, the error term is decomposed into a local and a spillover effect.

In a panel data setting, the $N \times 1$ error vector ε_t for each time period $t = 1, \ldots, T$, is expressed as:

$$\varepsilon_t = W_N \Psi_t + \xi_t \ , \tag{19.11}$$

where W_N is the weights matrix, ξ_t is a $N \times 1$ vector of local error components, and ψ_t is a $N \times 1$ vector of spillover error components. The two component vectors are assumed to consist of *i.i.d* terms, with respective variances σ_ψ^2 and σ_ξ^2, and are uncorrelated, $\mathrm{E}[\psi_{it} \xi_{jt}] = 0, \ \forall \ i, j, t$.

The resulting $N \times N$ cross-sectional error covariance matrix is then, for $t = 1, \ldots, T$:

$$\Omega_{t,N} = \mathrm{E}[\varepsilon_t \varepsilon_t'] = \sigma_\psi^2 W_N W_N' + \sigma_\xi^2 I_N. \tag{19.12}$$

In the homogeneous model, this is again unchanging across time periods, and the overall $NT \times NT$ error covariance matrix can be expressed as:

[14] For example, with W_N specified as first order contiguity, the spatial covariance in (19.10) only includes first and second order neighbors.

$$\Sigma_{NT} = \sigma_{\xi}^2 I_{NT} + \sigma_{\psi}^2 (I_T \otimes W_N W_N') \; .$$

Comparing (19.10 and 19.12), it can be readily seen that the range of covariance induced by the SEC model is a subset that of the SMA model, and hence also a case of *local* spatial externalities.

19.2.3.4 Common Factor Models

In the standard two-way error component regression model, each observational unit contains an unobserved error component due to individual heterogeneity and one due to a time period effect, in addition to the usual idiosyncratic error term (e.g., Baltagi, 2001, p. 31). In our notation:

$$\varepsilon_{it} = \mu_i + \lambda_t + u_{it} \; ,$$

with μ_i as the cross-sectional component, with variance σ_{μ}^2, λ_t as the time component, with variance σ_{λ}^2, and u_{it} as an idiosyncratic error term, assumed to be i.i.d with variance σ_u^2. The three random components are assumed to be zero mean and to be uncorrelated with each other. The random components μ_i are assumed to be uncorrelated across cross-sectional units, and the components λ_t are assumed to be uncorrelated across time periods. This model is standard, except that for our purposes, the data are stacked as cross-sections for different time periods. Consequently, the $N \times 1$ cross-sectional error vector ε_t for time period $t = 1, \ldots, T$, becomes:

$$\varepsilon_t = \mu + \lambda_t \iota_N + u_t, \tag{19.13}$$

where μ is a $N \times 1$ vector of cross-sectional error components μ_i, λ_t is a scalar time component, ι_N is a $N \times 1$ vector of ones, and u_t is a $N \times 1$ vector of idiosyncratic errors.

The structure in (19.13) results in a particular form of cross-sectional (spatial) correlation, due to the common time component:

$$\mathrm{E}[\varepsilon_t \varepsilon_t'] = \sigma_{\mu}^2 I_N + \sigma_{\lambda}^2 \iota_N \iota_N' + \sigma_u^2 I_N \; ,$$

where the subscript N indicates the dimension of the identity matrices. Note that the second term in this expression indicates equicorrelation in the cross-sectional dimension, i.e., the correlation between two cross-sectional units i, j equals σ_{λ}^2, no matter how far these units are apart. While perfectly valid as a model for general (global) cross-sectional correlation, this violates the distance decay effect that underlies spatial interaction theory.

The complete $NT \times 1$ error vector can be written as (see also Anselin, 1988a, p. 153):

$$\varepsilon = (\iota_T \otimes I_N)\mu + (I_T \otimes \iota_N)\lambda + u \; ,$$

where the subscripts indicate the dimensions, λ is a $T \times 1$ vector of time error components, u is a $NT \times 1$ vector of idiosyncratic errors, and the other notation is as before. The overal error variance covariance matrix then follows as:

$$\Sigma_{NT} = \sigma_\mu^2(\iota_T \iota_T' \otimes I_N) + \sigma_\lambda^2(I_T \otimes \iota_N \iota_N') + \sigma_u^2 I_{NT} \ .$$

Note the how the order of matrices in the Kronecker products differs from the standard textbook notation, due to the stacking by cross-section.

A recent extension of the error component model can be found in the literature on heterogeneous panels. Here, the time component is generalized and expressed in the form of an unobserved common effect or *factor* f_t to which all cross-sectional units are exposed. However, unlike the standard error component model, each cross-sectional unit has a distinct factor *loading* on this factor. The simplest form is the so-called one factor structure, where the error term is specified as:

$$\varepsilon_{it} = \delta_i f_t + u_{it} \ ,$$

with δ_i as the cross-sectional-specific loading on factor f_t, and u_{it} as an *i.i.d* zero mean error term. Consequently, cross-sectional (spatial) covariance between the errors at i and j follows from the the inclusion of the common factor f_t in both error terms:

$$\mathrm{E}[\varepsilon_{it}\varepsilon_{jt}] = \delta_i \delta_j \sigma_f^2 \ .$$

The common factor model has been extended to include multiple factors. In these specifications, a wide range of covariance structures can be expressed by including sufficient factors and through cross-sectional differences among the factor loadings (for further details, see Driscoll and Kraay, 1998; Pesaran, 2002; and Hsiao and Pesaran, 2008).

19.3 A Taxonomy of Spatial Panel Model Specifications

So far, we have considered the introduction of spatial effects for panel data in the form of spatial lag or spatial error models under extreme homogeneity. The point of departure was the pooled specification, (19.1), and lag and error models are obtained as outlined in Sects. 19.2.2 and 19.2.3. We now extend this taxonomy by introducing heterogeneity, both over time and across space, as well as by considering joint space-time dependence.

It should be noted that a large number of combinations of space-time heterogeneity and dependence are possible, although many of those suffer from identification problems and/or are not estimable in practice. In our classification here, we purposely limit the discussion to models that have seen some empirical applications (other, more extensive typologies can be found in Anselin, 1988a, Chap. 4; Anselin, 2001b; Elhorst, 2001, 2003).

19.3.1 Temporal Heterogeneity

19.3.1.1 General Case

Temporal heterogeneity is introduced in the familiar way in fixed effects models, by allowing time-specific intercepts and/or slopes, and in random effects models, by incorporating a random time component or factor (see Sect. 19.2.3.4). The addition of a spatially lagged dependent variable or spatially correlated error term in these models is straightforward. For example, consider a pooled model with time-specific intercept and slope coefficient to which a spatially autoregressive error term is added. The cross-section in each period $t = 1, \ldots, T$, is:

$$y_t = \alpha_t + X_t \beta_t + \varepsilon_t, \tag{19.14}$$

with

$$\varepsilon_t = \theta_t W_N \varepsilon_t + u_t$$

where θ_t is a period-specific spatial autoregressive parameter, α_t is the period-specific intercept and β_t a $(K-1) \times 1$ vector of period-specific slopes. Since T is fixed (and the asymptotics are based on $N \rightarrow \infty$), this model is a straightforward replication of T cross-sectional models. A spatial lag specification is obtained in a similar way.

19.3.1.2 Spatial Seemingly Unrelated Regressions

A generalization of the fixed effects model that has received some attention in the empirical literature (e.g., Rey and Montouri, 1999) allows the cross-sectional error terms ε_t to be correlated over time periods. This imposes very little structure on the form of the temporal dependence and is the spatial counterpart of the classic SURE model. It is referred to as the *spatial* SUR model (see Anselin, 1988a, Chap. 10; and Anselin, 1988b). In matrix form, the equation for the cross-sectional regression in each time period $t = 1, \ldots, T$, is as in (19.14), but now with the constant term included in the vector β_t:

$$y_t = X_t \beta_t + \varepsilon_t, \tag{19.15}$$

with the cross-equation (temporal) correlation in general form, as:

$$\mathrm{E}[\varepsilon_t \varepsilon_s'] = \sigma_{ts} I_N, s \neq t \ ,$$

where σ_{ts} is the temporal covariance between s and t (by convention, the variance terms are expressed as σ_t^2). In stacked form (T cross-sections), the model is:

$$y = X\beta + \varepsilon, \tag{19.16}$$

with

$$\mathrm{E}[\varepsilon\varepsilon'] = \Sigma_T \otimes I_N \tag{19.17}$$

and Σ_T is the $T \times T$ temporal covariance matrix with elements σ_{ts}.

Spatial correlation can be introduced as a spatial lag specification or a spatial error specification. Consider the spatial lag model first (see Anselin, 1988a for details). In each cross-section (with $t = 1, \ldots, T$), the standard spatial lag specification holds, but now with a time-specific spatial autoregressive coefficient ρ_t:

$$y_t = \rho_t W_N y_t + X_t \beta_t + \varepsilon_t \ .$$

To consider the full system, let β be a $TK \times 1$ vector of the stacked time-specific β_t, for $t = 1, \ldots, T$.[15] The corresponding $NT \times KT$ matrix X of observations on the explanatory variables then takes the form:

$$X = \begin{pmatrix} X_1 & 0 & \ldots & 0 \\ 0 & X_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & X_T \end{pmatrix}. \tag{19.18}$$

Also, let the spatial autoregressive coefficients be grouped in a $T \times T$ diagonal matrix R_T, as:

$$R_T = \begin{pmatrix} \rho_1 & 0 & \ldots & 0 \\ 0 & \rho_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \rho_T \end{pmatrix}.$$

The full system can then be expressed concisely as:

$$y = (R_T \otimes W_N) y + X\beta + \varepsilon, \tag{19.19}$$

with the error covariance matrix as in (19.17).

In empirical practice, besides the standard hypothesis tests on diagonality of the error covariance matrix and stability of the regression coefficients over time, interest in the *spatial lag* SUR model will center on testing the hypothesis of homogeneity of the spatial autoregressive coefficients, or, $\mathrm{H}_0 : \rho_1 = \rho_2 = \ldots = \rho_T = \rho$. If this null hypothesis can be maintained, a simplified model can be implemented:

$$y = \rho(I_T \otimes W_N) y + X\beta + \varepsilon \ .$$

Spatial error autocorrelation can be introduced in the spatial SUR model in the form of a SAR or SMA process for the error terms (see Anselin, 1988a, Chap. 10). For example, consider the following SAR error process for the cross-section in each time period $t = 1, \ldots, T$:

$$\varepsilon_t = \theta_t W_N \varepsilon_t + u_t \ . \tag{19.20}$$

[15] We are assuming the same number of explanatory variables (K) in each equation, but this can be readily generalized to allow the number of explanatory variables to vary by time period.

The cross-equation covariance is introduced through the remainder error term u_t, for which it is assumed that $\mathrm{E}[u_t] = 0$, $\mathrm{E}[u_t u_t'] = \sigma_t^2 I_N$, and $\mathrm{E}[u_t u_s'] = \sigma_{ts} I_N$, for $t \neq s$. As a result, the covariance matrix for the stacked $NT \times 1$ error vector u becomes the counterpart of (19.17):

$$\mathrm{E}[uu'] = \Sigma_T \otimes I_N \ ,$$

with, as before, Σ_T as a $T \times T$ matrix with elements σ_{ts}.

The SAR error process in (19.20) can also be written as:

$$\varepsilon_t = (I_N - \theta_t W_N)^{-1} u_t \ ,$$

or, using the simplifying notation $B_{t,N} = (I_N - \theta_t W_N)$, as:

$$\varepsilon_t = B_{t,N}^{-1} u_t \ .$$

The overall cross-equation covariance between error vectors ε_t and ε_s then becomes:

$$\mathrm{E}[\varepsilon_t \varepsilon_s'] = B_{t,N}^{-1} \, \mathrm{E}[u_t u_s'] B_{s,N}^{-1'} = \sigma_{ts} \, B_{t,N}^{-1} \, B_{s,N}^{-1'} \ ,$$

which illustrates how the simple SUR structure induces space-time covariance as well (the $B_{t,N}^{-1}$ matrices are not diagonal).

In stacked form, the error process for the $NT \times 1$ error vector ε can be written as:

$$\varepsilon = B_{NT}^{-1} u \ ,$$

with B_{NT} as the matrix:

$$B_{NT} = [I_{NT} - (\Theta_T \otimes W_N)], \tag{19.21}$$

and Θ_T as a $T \times T$ diagonal matrix containing the spatial autoregressive coefficients $\theta_t, t = 1, \ldots, T$. The overall error covariance matrix for the stacked equations then becomes:

$$\mathrm{E}[\varepsilon\varepsilon'] = B_{NT}^{-1}(\Sigma_T \otimes I_N) B_{NT}^{-1'}. \tag{19.22}$$

As in the spatial lag SUR model, specific interest in the spatial error SUR model centers on the homogeneity of the spatial autoregressive parameters, $\mathrm{H}_0 : \theta_1 = \theta_2 = \ldots = \theta_T = \theta$. If the homogeneity holds, the expression for B_{NT} (19.21) simplifies to:

$$B_{NT} = [I_T \otimes (I_N - \theta W_N)] \ .$$

19.3.2 Spatial Heterogeneity

We limit our attention in the treatment of spatial heterogeneity to models with unobserved heterogeneity, specified in the usual manner as either fixed effects or random effects. Both have been extended with spatial lag and spatial error specifications.

19.3.2.1 Fixed Effects Models

The classic fixed effects model (e.g., Baltagi, 2001, pp. 12–15; and Arellano, 2003, pp. 11–18) includes an individual specific "dummy variable" to capture unobserved heterogeneity. For each observation i,t this yields, keeping the same notation as before:

$$y_{i,t} = \alpha_i + x_{it}\beta + \varepsilon_{it}$$

for $i = 1,\ldots,N$, $t = 1,\ldots,T$, and with an additional constraint of the form $\sum_i \alpha_i = 0$, such that the individual effects α_i are separately identifiable from the constant term in β.

As is well known, consistent estimation of the individual fixed effects is not possible when $N \rightarrow \infty$, due to the incidental parameter problem. Since spatial models rely on the asymptotics in the cross-sectional dimension to obtain consistency and asymptotic normality of estimators, this would preclude the fixed effects model from being extended with a spatial lag or spatial error term (Anselin, 2001b).

Nevertheless, it has been argued that when the interest is primarily in obtaining consistent estimates for the β coefficients, the use of *demeaned* spatial regression models may be appropriate, for example, using the standard maximum likelihood estimation expressions (Elhorst, 2003, p. 250–251).

There are a number of aspects of this approach that warrant closer attention. One is that the demeaning operator takes on a different form from the usual expression in the literature, since the observations are stacked as cross-sections for different time periods. Also, the demeaned models no longer contain a constant term, which may be incompatible with assumptions made by standard spatial econometric software. More importantly, the variance covariance matrix of the demeaned error terms is no longer $\sigma^2_\varepsilon I$, but becomes $\sigma^2_\varepsilon Q$, where Q is the demeaning operator (this aspect is ignored in the likelihood functions presented in Elhorst, 2003, p. 250).

To illustrate these points, consider a fixed effects spatial lag model in stacked form, using the same setup as in (19.5), with the addition of the fixed effects:

$$y = \rho(I_T \otimes W_N)y + (\iota_T \otimes \alpha) + X\beta + \varepsilon, \tag{19.23}$$

where α is a $N \times 1$ vector of individual fixed effects, with the constraint that $\alpha'\iota_N = 0$, and, as before, $\mathrm{E}[\varepsilon\varepsilon'] = \sigma^2_\varepsilon I_{NT}$. Note the difference with the classic formulation in the Kronecker product for the fixed effects, due to the stacking of cross-sections, rather than individual time series.

The demeaned form of (19.23) is obtained by substracting the average for each cross-sectional unit computed over the time dimension, which wipes out the individual fixed effects (as well as the constant term). Formally, this can be expressed as:

$$Q_{NT}y = \rho(I_T \otimes W_N)Q_{NT}y + Q_{NT}X\beta + Q_{NT}\varepsilon, \tag{19.24}$$

where Q_{NT} is the demeaning operator (and $Q_{NT}X$ and β no longer contain a constant term). The demeaning operator is a $NT \times NT$ matrix that takes the form:

$$Q_{NT} = I_{NT} - (\iota_T \iota_T' / T \otimes I_N)$$

with, as before, ι as a vector of ones and the subscripts denoting the dimension of vectors and matrices. Again, note the difference with the standard textbook notation, due to stacking by cross-section. The matrix Q_{NT} is idempotent, and, as a result, the variance of the error in (19.24) becomes:

$$\mathrm{E}[\varepsilon\varepsilon'] = \sigma_\varepsilon^2 Q_{NT} \ ,$$

where Q_{NT} is singular. Consequently, $|Q_{NT}| = 0$ and the regular inverse of the matrix Q_{NT} does not exist. In the non-spatial case, this problem disappears because of the properties of the generalized inverse Q^- (see, e.g., Hsiao, 1986, p. 222). In the absence of spatial variables, the regression in demeaned X and y can be treated as a special case of GLS estimation, with Q^- as the generalized inverse, such that $QQ^-Q = Q$. As a result $\hat{\beta} = (X'Q'Q^-QX)^{-1}X'Q'Q^-Qy = (X'QX)^{-1}X'Qy$, which no longer involves the generalized inverse in the actual calculations.

However, the log-likelihood for the spatial lag model with demeaned variables is based on the multivariate normality of the error term ε. In this case, the singularity of Q_{NT} constitutes a problem since the (joint unconditional) likelihood becomes degenerate.[16] In the non-spatial literature on dynamic panels with individual fixed effects, this problem is avoided by considering a transformed likelihood function based on first differences of the variables (see Hsiao et al., 2002). The extension of this idea to spatial models remains to be considered.

A likelihood approach to the spatial error model faces a similar complication.

19.3.2.2 Random Effects Models

In the random effects approach to modeling unobserved heterogeneity, interest has centered on incorporating spatial error correlation into the regression error term, in addition to the standard cross-sectional random component. Note that the latter induces serial correlation over time (of the equi-correlated type). Here, we focus attention on the one-way error component specification (e.g, Baltagi, 2001, pp. 15–20; Arellano, 2003, Chap. 3).[17]

In contrast to the fixed effects case, asymptotics along the cross-sectional dimension (with $N \rightarrow \infty$) present no problem for random effects models. The standard specification of the error term in this model, is, for each i,t:

$$\varepsilon_{it} = \mu_i + \nu_{it} \ ,$$

[16] In Elhorst, 2003 (p. 250), the log-likelihood uses $\sigma^2 I$ as the error variance, not $\sigma^2 Q_{NT}$.

[17] Explicit space-time dependence is treated in Sect. 19.3.3.

where $\mu_i \backsim \text{IID}(0, \sigma_\mu^2)$ is the cross-sectional random component, and $\nu_{it} \backsim \text{IID}(0, \sigma_\nu^2)$ is an idiosyncratic error term, with μ_i and ν_{it} independent from each other. In each cross-section, for $t = 1, \ldots, T$, the $N \times 1$ error vector ε_t becomes:

$$\varepsilon_t = \mu + \nu_t, \tag{19.25}$$

where μ is a $N \times 1$ vector of cross-sectional random components.

Spatial error autocorrelation can be introduced into this expression in a number of different ways. A first approach follows the analogue from the time domain and specifies a SAR process for the error component ν_t, for $t = 1, \ldots, T$ (Anselin, 1988a, p. 153; and, more recently, Baltagi et al., 2003):

$$\nu_t = \theta W_N \nu_t + u_t, \tag{19.26}$$

with θ as the spatial autoregressive parameter (constant over time), W_N as the spatial weights matrix, and u_t as an i.i.d idiosyncratic error term with variance σ_u^2.

Using the notation $B_N = I_N - \theta W_N$, we obtain the familiar result:

$$\nu_t = (I_N - \theta W_N)^{-1} u_t = B_N^{-1} u_t \ .$$

This illustrates how the spatial autocorrelation pertains to a cross-section in each time period (the error vector u_t) separately. In stacked form, the $NT \times 1$ error term then becomes:

$$\varepsilon = (\iota_T \otimes I_N)\mu + (I_T \otimes B_N^{-1})u, \tag{19.27}$$

where $u \backsim \text{IID}(0, \sigma_u^2 I_{NT})$ is a $NT \times 1$ vector of idiosyncratic errors. The corresponding variance–covariance matrix for ε follows as:

$$\Sigma_{NT} = \text{E}[\varepsilon\varepsilon'] = \sigma_\mu^2(\iota_T \iota_T' \otimes I_N) + \sigma_u^2[I_T \otimes (B_N' B_N)^{-1}] \ . \tag{19.28}$$

Note that the first component induces correlation in the time dimension, but not in the cross-sectional dimension, whereas the opposite holds for the second component (correlation only in the cross-sectional dimension).

A second specification for spatial correlation in this model applies the SAR process first and the error components specification to its remainder error (Kapoor et al., 2007). Consider a SAR process for the $NT \times 1$ error vector ε:

$$\varepsilon = \theta(I_T \otimes W_N)\varepsilon + \nu \ ,$$

or, using similar notation as the spatially correlated component in (19.27):

$$\varepsilon = (I_T \otimes B_N^{-1})\nu \ .$$

Now, the innovation vector ν is specified as a one way error component model (Kapoor et al., 2007):

$$\nu = (\iota_T \otimes I_N)\mu + u \ ,$$

with μ as the $N \times 1$ vector of cross-sectional random components, and $u \backsim \text{IID}(0, \sigma_u^2 I_{NT})$. In stacked form, the full error vector follows as:

$$\varepsilon = (I_T \otimes B^{-1})[(\iota_T \otimes I_N)\mu + u] \; .$$

The corresponding error variance covariance matrix is:

$$\Sigma_{NT} = \text{E}[\varepsilon\varepsilon'] = (I_T \otimes B_N^{-1})[\sigma_\mu^2(\iota_T \iota_T' \otimes I_N) + \sigma_u^2 I_{NT}](I_T \otimes B_N^{-1'}). \tag{19.29}$$

Again, this model combines both time-wise as well as cross-sectional correlation, with the latter pertaining to both the time specific error u_t as well as the time invariant error component μ.

Recently, an encompassing specification was suggested (Baltagi et al., 2006) that includes both forms as special cases and provides a useful starting point for a wide range of specification tests. In this model, an explicit distinction is made between *permanent* and *time variant* spatial correlation. The former introduces a spatial autoregressive process for the $N \times 1$ error component μ:

$$\mu = \theta_1 W_N \mu + u_1 \; ,$$

or, with $A = I_N - \theta_1 W_N$, $\mu = A^{-1} u_1$, where u_1 is a $N \times 1$ vector of time invariant idiosyncratic errors with variance σ_{u1}^2.

Time variant spatial correlation is included as a spatial autoregressive process for each time-specific error vector ν_t, as:

$$\nu_t = \theta_2 W_N \nu_t + u_{2t} \; ,$$

or, with $B = I_N - \theta_2 W_N$, $\nu_t = B^{-1} u_{2t}$, where u_{2t} is a $N \times 1$ vector of time invariant idiosyncratic errors with variance σ_{u2}^2.

The stacked error term then follows as:

$$\varepsilon = (\iota_T \otimes I_N) A^{-1} u_1 + (I_T \otimes B^{-1}) u_2 \; ,$$

and u_2 consists of the stacked u_{2t}.

The corresponding error variance–covariance matrix follows as:

$$\Sigma = \sigma_{u1}^2 [\iota_T \iota_T' \otimes (A'A)^{-1}] + \sigma_{u2}^2 [I_T \otimes (B'B)^{-1}] \; .$$

The first element in this expression contains both time and spatial correlation, whereas the second term only contains spatial correlation.

The two earlier specifications are found by imposing parameter constraints on the encompassing form. More precisely, for $\theta_1 = 0$, the first model is obtained, and for

$\theta_1 = \theta_2$, the second model follows. When both $\theta_1 = \theta_2 = 0$, the standard non-spatial random effects model is obtained.

Statistical inference for error components models with spatial SAR processes can be carried out as a special case of models with non-spherical error covariance. This is addressed in Sects. 19.4 and 19.5.

19.3.3 Spatio-Temporal Models

The incorporation of dependence in both time and space dimensions in an econometric specification adds an additional order of difficulty to the identification of the $NT \times (NT-1)/2$ elements of the variance covariance matrix. An important concept in this regard is the notion of *separability*. Separability requires that a $NT \times NT$ space-time covariance matrix Σ_{NT} can be decomposed into a component due to space and a component due to time (see, e.g., Mardia and Goodall, 1993), or:

$$\Sigma_{NT} = \Sigma_T \otimes \Sigma_N \ ,$$

where Σ_T is a $T \times T$ variance covariance matrix for the time-wise dependence and Σ_N is a $N \times N$ variance covariance matrix for the spatial dependence.[18] This ensures that the space-time dependence declines in a multiplicative fashion over the two dimensions. It also addresses a central difficulty in space-time modeling, i.e., the lack of a common "distance" metric that works both in the cross-sectional and the time dimension. The approach taken in spatial panel econometrics is to define "neighbors" in space by means of a spatial weights matrix and "neighbors" in time by means of the customary time lags. However, the speed of the dynamic space-time process may not be compatible with these choices, leading to further misspecification.

19.3.3.1 Model Taxonomy

Ignoring for now any space-time dependence in the error terms, we can distinguish four basic forms to introduce correlation in both space and time in panel data models (following Anselin, 2001b, p. 317–318). As before, we focus on models where $N \gg T$ and do not consider specifications where the time dimension is an important aspect of the model.[19] To facilitate exposition, we express these models for a $N \times 1$ cross-section at time $t = 1, \ldots, T$.

[18] The notion of separable stationary spatio-temporal processes originates in the geostatistical literature, but can be readily applied to the current framework. Extension to non-separable structures have been suggested in the recent literature (e.g., Cressie and Huang, 1999).

[19] In the statistical literature, specifications of space-time dependence are often conceptualized as hierarchical or multilevel models. This is beyond the scope of our current review (see, for example, Waller et al., 1997a, b; Wikle et al., 1998; Banerjee et al., 2004, and the extensive set of references therein).

Pure space recursive models, in which the dependence pertains only to neighboring locations in a previous period:

$$y_t = \gamma W_N y_{t-1} + X_t \beta + \varepsilon_t, \tag{19.30}$$

with γ as the space-time autoregressive parameter, and $W_N y_{t-1}$ as a $N \times 1$ vector of observations on the spatially lagged dependent variable at $t-1$. Note that this can be readily extended with time and spatial lags of the explanatory variables, X_{t-1} or $W_N X_t$. However, since $W_N y_{t-1}$ already includes $W_N X_{t-1}$, adding a term of this form would create identification problems. This is sometimes overlooked in other taxonomies of dynamic space-time models (e.g., in the work of Elhorst, 2001, p. 121, where space-time lags for both the dependent and the exploratory variables are included in the specification).

Consider the space-time *multiplier* more closely. Start by substituting the equation for y_{t-1} in (19.30), which yields:

$$y_t = \gamma W_N [\gamma W_N y_{t-2} + X_{t-1} \beta + \varepsilon_{t-1}] + X_t \beta + \varepsilon_t \ ,$$

or,

$$y_t = \gamma^2 W_N^2 y_{t-2} + X_t \beta + \gamma W_N X_{t-1} \beta + \varepsilon_t + \gamma W_N \varepsilon_{t-1} \ .$$

Successive substitution reveals a space-time multiplier that follows from a series of consecutively higher orders of both spatial and time lags applied to the X (and error terms). Also, since the spatial dependence takes one period to manifest itself, this specification becomes quite suitable to study spatial diffusion phenomena (see the early discussion in Upton and Fingleton, 1985; and Dubin, 1995).

Time-space recursive models, in which the dependence relates to both the location itself as well as its neighbors in the previous period:

$$y_t = \phi y_{t-1} + \gamma W_N y_{t-1} + X_t \beta + \varepsilon_t \ , \tag{19.31}$$

with ϕ as the serial (time) autoregressive parameter, operating on the cross-section of dependent variables at $t-1$. Spatially lagged contemporaneous explanatory variables ($W_N X_t$) may be included as well, but time lagged explanatory variables will result in identification problems. This model has particular appeal in space-time forecasting (e.g., Giacomini and Granger, 2004).

Again, the nature of the space-time multiplier can be assessed by substituting the explicit form for the spatially and time lagged terms:

$$\begin{aligned} y_t = \ & \phi [\phi y_{t-2} + \gamma W_N y_{t-2} + X_{t-1} \beta + \varepsilon_{t-1}] \\ & + \gamma W_N [\phi y_{t-2} + \gamma W_N y_{t-2} + X_{t-1} \beta + \varepsilon_{t-1}] \\ & + X_t \beta + \varepsilon_t \ , \end{aligned}$$

or,

$$\begin{aligned} y_t = &(\phi^2 + 2\phi\gamma W_N + \gamma^2 W_N^2) y_{t-2} \\ &+ X_t\beta + (\phi + \gamma W_N) X_{t-1}\beta \\ &+ \varepsilon_t + (\phi + \gamma W_N)\varepsilon_{t-1} \ , \end{aligned}$$

revealing a much more complex form for the effect of space-time lagged explanatory variables (and errors), including the location itself as well as its neighbors.

Time-space simultaneous models, which include a time lag for the location itself together with a contemporaneous spatial lag:

$$y_t = \phi y_{t-1} + \rho W_N y_t + X_t\beta + \varepsilon_t \ ,$$

with ρ as the (contemporaneous) spatial autoregressive parameter.

The mulitplier in this model is complex, due to the combined effect of the cross-sectional *spatial* multiplier (in each period) and the space-time multiplier that follows from the time lag in the dependent variable. First, consider the pure cross-sectional multiplier:

$$y_t = (I_N - \rho W_N)^{-1}[\phi y_{t-1} + X_t\beta + \varepsilon_t] \ .$$

Next, substitute the corresponding expression for y_{t-1}:

$$\begin{aligned} y_t = &(I_N - \rho W_N)^{-1}[\phi[(I_N - \rho W_N)^{-1}(\phi y_{t-2} \\ &+ X_{t-1}\beta + \varepsilon_{t-1})] + X_t\beta + \varepsilon_t] \ , \end{aligned}$$

which yields:

$$\begin{aligned} y_t = &\phi^2 (I_N - \rho W_N)^{-2} y_{t-2} \\ &+ (I_N - \rho W_N)^{-1} X_t\beta + \phi (I_N - \rho W_N)^{-2} X_{t-1}\beta \\ &+ (I_N - \rho W_N)^{-1}\varepsilon_t + \phi (I_N - \rho W_N)^{-2}\varepsilon_{t-1} \ . \end{aligned}$$

From this it follows that the inclusion of any spatially lagged X in the original specification will lead to identification problems.

Time-space dynamic models, where all three forms of lags for the dependent variable are included:

$$y_t = \phi y_{t-1} + \rho W_N y_t + \gamma W_N y_{t-1} + X_t\beta + \varepsilon_t \ .$$

While this model is sometimes suggested as a *general* space-time specification, it results in complex nonlinear constraints on the parameters, and, in practice, often suffers from identification problems. For example, focusing only on the time lagged terms and substituting their expression for $t-1$ (and rearranging terms) yields:

$$
\begin{aligned}
y_t = \; & \phi[\phi y_{t-2} + \rho W_N y_{t-1} + \gamma W_N y_{t-2} + X_{t-1}\beta + \varepsilon_{t-1}] \\
& + \gamma W_N[\phi y_{t-2} + \rho W_N y_{t-1} + \gamma W_N y_{t-2} + X_{t-1}\beta + \varepsilon_{t-1}] \\
& + \rho W_N y_t + X_t\beta + \varepsilon_t \; ,
\end{aligned}
$$

or, grouping by time period:

$$
\begin{aligned}
y_t = \; & \rho W_N y_t + X_t\beta + \varepsilon_t \\
& + \phi\rho W_N y_{t-1} + \gamma\rho W_N^2 y_{t-1} + \phi X_{t-1}\beta + \gamma W_N X_{t-1}\beta \\
& + \phi\varepsilon_{t-1} + \gamma W_N \varepsilon_{t-1} \\
& + \phi^2 y_{t-2} + \gamma\phi W_N y_{t-2} + \gamma^2 W_N^2 y_{t-2} \; .
\end{aligned}
$$

The same types of space-time dependence processes can also be specified for the error terms in panel data models (e.g., Fazekas et al., 1994). However, combinations of both spatially lagged dependent variables and spatially lagged error terms may lead to identification problems unless the parameters of the explanatory variables are non-zero. An alternative form of error space-time dependence takes the error components approach, to which we turn briefly.

19.3.3.2 Error Components with Space-Time Dependence

The starting point for including explicit serial dependence (in addition to the equicorrelated form) in random effects models is the spatially autocorrelated form considered in (19.25–19.26). However, instead of the indiosyncratic error u_t in 19.26, a serially correlated term ζ_t is introduced (Baltagi et al., 2007):

$$
\nu_t = \theta W_N \nu_t + \zeta_t \tag{19.32}
$$

with

$$
\zeta_t = \phi\,\zeta_{t-1} + u_t, \tag{19.33}
$$

where, as before, u_t is used to denote the idiosyncratic error, and $t = 1, \ldots, T$. The counterpart of the $N \times 1$ cross-sectional error vector ε_t in (19.27) becomes:

$$
\varepsilon_t = (I_N - \theta W_N)^{-1}\zeta_t = B_N^{-1}\zeta_t \; ,
$$

with ζ replacing the original error u. In stacked form, this becomes:

$$
\varepsilon = (\iota_T \otimes I_N)\mu + (I_T \otimes B_N^{-1})\zeta \; ,
$$

with μ of dimension $N \times 1$ and both ε and ζ of dimension $NT \times 1$. The serial correlation in ζ will yield serial covariances of the familiar AR(1) form, with:

$$
\mathrm{E}[\zeta_{i,t}\zeta_{i,t-k}] = \sigma_u^2 \left(\frac{\phi^k}{1-\phi^2}\right) \; ,
$$

for $k = 0, \ldots, T-1$, and $i = 1, \ldots, N$, where σ_u^2 is the variance of the error term u. Grouping these serial covariances into a $T \times T$ variance covariance matrix Ω_T yields the overall variance covariance matrix for ε as (Baltagi et al., 2007):

$$\Sigma_{NT} = \mathrm{E}[\varepsilon\varepsilon'] = \sigma_\mu^2(\iota_T\iota_T' \otimes I_N) + [\Omega_T \otimes (B_N'B_N)^{-1}] \, .$$

19.4 Estimation of Spatial Panel Models

The estimation of panel data models that include spatially lagged dependent variables and/or spatially correlated error terms follows as a direct extension of the theory developed for the single cross-section. In the first case, the endogeneity of the spatial lag must be dealt with, in the second, the non-spherical nature of the error variance covariance matrix must be accounted for. Two main approaches have been suggested in the literature, one based on the maximum likelihood principle, the other on method of moments techniques. We consider each in turn.

We limit our attention to models with a parameterized form for the spatial dependence, specified as a spatial autoregressive process.[20] Note that some recent results in the panel econometrics literature have also addressed estimation in models with general, unspecified cross-sectional correlation (see, e.g., Driscoll and Kraay, 1998; Coakley et al., 2002; Pesaran, 2002).

19.4.1 Maximum Likelihood Estimation

The theoretical framework for maximum likelihood estimation of spatial models in the single cross-section setup is by now well developed (see, among others, Ord, 1975; Mardia and Marshall, 1984; Anselin, 1988a; Cressie, 1993; Anselin and Bera, 1998). While the regularity conditions are non-standard, and require a consideration of triangular arrays (Kelejian and Prucha, 1999), the results for error terms with a Gaussian distribution are fairly well established.

In practice, estimation consists of applying a non-linear optimization to the log-likelihood function, which (in most circumstances) yields a consistent estimator from the numerical solution to the first order conditions. Asymptotic inference is based on asymptotic normality, with the asymptotic variance matrix derived from the information matrix. This requires the second order partial derivatives of the log-likelihood, for which analytical solutions exist in many of the models considered (for technical details, see the review in Anselin and Bera, 1998).

A main obstacle in the practical implementation of ML estimation in a single cross-section is the need to compute a Jacobian determinant for an N-dimensional matrix (the dimension of the cross-section). In panel data models, this Jacobian is

[20] Models with other forms for the error dependence have seen limited attention in a panel data context and are not considered here.

of dimension $N \times T$, but it can often be simplified to a product of T N-dimensional determinants. The classic solution to this problem is to decompose the Jacobian in terms of the eigenvalues of the spatial weights matrix. For example, in the spatial lag model, the Jacobian would be $|I_N - \rho W_N| = \prod_i (1 - \rho \omega_i)$, with ω_i as the eigenvalues of W_N (Ord, 1975).[21] For large cross-sections, the computation of the eigenvalues becomes numerically unstable, precluding this method from being applicable. Alternative solutions avoid the computation of the Jacobian determinant, but instead approximate it by a polynomial function or by means of simulation methods (Barry and Pace, 1999). Other methods are based on Cholesky or LU decomposition methods that exploit the sparsity of the spatial weights (Pace and Barry, 1997), or use a characteristic polynomial approach (Smirnov and Anselin, 2001).

We now briefly review a number of useful log-likelihood expressions that result when incorporating spatial lag or spatial error terms in panel data settings. Numerical procedures to carry out estimation and inference can be implemented along the same lines as for the single cross-section, and will not be further elaborated.

19.4.1.1 Spatial Lag Models

As a point of departure, consider the pooled spatial lag model given in (19.5). Assuming a Gaussian distribution for the error term, with $\varepsilon \backsim N(0, \sigma_\varepsilon^2 I_{NT})$, the log-likelihood (ignoring the constants) follows as:

$$L = \ln |I_T \otimes (I_N - \rho W_N)| - \frac{NT}{2} \ln \sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2} \varepsilon' \varepsilon \ ,$$

with $\varepsilon = y - \rho (I_T \otimes W_N) y - X\beta$, and $|I_T \otimes (I_N - \rho W_N)|$ as the Jacobian determinant of the spatial transformation. Given the block diagonal structure of the Jacobian, the log-likelihood further simplifies to:

$$L = T \ln |I_N - \rho W_N| - \frac{NT}{2} \ln \sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2} \varepsilon' \varepsilon, \tag{19.34}$$

which boils down to a repetition of the standard cross-sectional model in T cross-sections.

Generalizing this model slightly, we now assume $\varepsilon \backsim N(0, \Sigma)$ to allow for more complex error covariance structures (including spatial correlation). The log-likelihood remains essentially the same, except for the new error covariance term:

$$L = T \ln |I_N - \rho W_N| - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \varepsilon' \Sigma^{-1} \varepsilon. \tag{19.35}$$

[21] In practice, the log Jacobian is used, with $\ln |I_N - \rho W_N| = \sum_i \ln (1 - \rho \omega_i)$.

Two special cases result by introducing some structure into the variance covariance matrix Σ. First, consider the classic one-way error components model from (19.25), which, in stacked form, becomes (again, using cross-sections for T time periods and assuming a Gaussian distribution for ε):

$$\varepsilon = (\iota_T \otimes I_N)\mu + u \, .$$

The error covariance matrix follows as:

$$\Sigma_{NT} = \mathrm{E}[\varepsilon\varepsilon'] = \sigma_\mu^2(\iota_T\iota_T' \otimes I_N) + \sigma_u^2 I_{NT} \, . \tag{19.36}$$

Using standard results, the inverse and determinant of this $NT \times NT$ matrix can be expressed in terms of matrix determinants and inverses of orders N and T only. Inserting (19.36 into 19.35) yields the log-likelihood for the spatial lag model with error components as:

$$\begin{aligned} L = {} & T \ln |I_N - \rho W_N| - \tfrac{1}{2} \ln |\sigma_\mu^2(\iota_T\iota_T' \otimes I_N) + \sigma_u^2 I_{NT}| \\ & - \tfrac{1}{2}\varepsilon' \left[\sigma_\mu^2(\iota_T\iota_T' \otimes I_N) + \sigma_u^2 I_{NT}\right]^{-1} \varepsilon \, . \end{aligned}$$

A second specification of interest is the SUR model that includes a spatial lag term, (19.19). Its log-likelihood can be obtained in a similar fashion. Using the same notation and stacking of observation matrices and parameters as in (19.18–19.19), the log Jacobian follows as $\ln |I_{NT} - (R_T \otimes W_N)|$. The block diagonal structure of the matrix can be exploited to simplify this expression to $\sum_t \ln |I_N - \rho_t W_N|$ (with the sum over $t = 1, \ldots, T$). Using (19.17) for the error variance covariance matrix in the SUR model, the log-likelihood follows as:

$$L = \sum_t \ln |I_N - \rho_t W_N| - \frac{N}{2} \ln |\Sigma_T| - \frac{1}{2}\varepsilon' \left(\Sigma_T^{-1} \otimes I_N\right) \varepsilon \, ,$$

with $\varepsilon = [I_{NT} - (R_T \otimes W_N)]y - X\beta$ (for further details, see Anselin, 1988a, pp. 145–146).

19.4.1.2 Spatial Error Models

The log-likelihood functions for the various spatial error models considered in this chapter follow directly as special cases of the standard result for maximum likelihood estimation with a non-spherical error covariance (Magnus, 1978). With $\varepsilon \backsim N(0, \Sigma)$ as the error vector, the familiar expression for the log-likelihood is (ignoring the constant terms):

$$L = -\frac{1}{2} \ln |\Sigma| - \frac{1}{2}\varepsilon'\Sigma^{-1}\varepsilon \, .$$

In the pooled model with SAR error terms, (19.9), the relevant determinant and inverse matrix are:

$$|I_T \otimes (B_N' B_N)^{-1}| = |B_N|^{-2T}$$

with B_N as in (19.9), and:

$$\Sigma_{NT}^{-1} = \frac{1}{\sigma_u^2}[I_T \otimes (B_N' B_N)] \, .$$

The corresponding log-likelihood function is then:

$$L = -\frac{NT}{2} \ln \sigma_u^2 + T \ln |B_N| \\ -\frac{1}{2\sigma_u^2} \varepsilon' [I_T \otimes (B_N' B_N)] \varepsilon \, ,$$

with $\varepsilon = y - X\beta$. The estimates for the regression coefficient β are the result of a spatial FGLS, using a consistent estimator for θ:

$$\hat{\beta} = [X'(I_T \otimes B_N' B_N) X]^{-1} X'(I_T \otimes B_N' B_N) y. \tag{19.37}$$

Exploiting the block diagonal nature of $B_N' B_N$, this is equivalent to a regression of the stacked spatially filtered dependent variables, $(I_N - \theta W_N) y_t$ on the spatially filtered explanatory variables $(I_N - \theta W_N) X_t$, as a direct generalization of the single cross-section case.

Two special cases are of particular interest. One is the random effects model with spatial error correlation. Its error variance covariance matrix, (19.28), can be simplified in order to facilitate the computation of the determinant and inverse term needed in the log-likelihood. Set $\eta = \sigma_\mu^2 / \sigma_u^2$, such that $\Sigma_{NT} = \sigma_u^2 \Psi_{NT}$, with:

$$\Psi_{NT} = \iota_T \iota_T' \otimes \eta I_N + [I_T \otimes (B_N' B_N)^{-1}] \, ,$$

using the same notation and observation stacking as for (19.28). This particular expression allows the determinant and inverse to be obtained as (see Anselin, 1988a, p. 154, for details):

$$|\Psi_{NT}| = |(B_N' B_N)^{-1} + (T\eta) I_N| |B_N|^{-2(T-1)}$$

and,

$$\Psi_{NT}^{-1} = \frac{\iota_T \iota_T'}{T} \otimes [(B_N' B_N)^{-1} + (T\eta) I_N]^{-1} + (I_T - \frac{\iota_T \iota_T'}{T}) \otimes (B_N' B_N) \, .$$

The log-likelihood thus becomes:

$$L = -\frac{NT}{2}\ln\sigma_u^2 - (T-1)\ln|B_N|$$
$$-\frac{1}{2}\ln|(B_N'B_N)^{-1} + (T\eta)I_N|$$
$$-\frac{1}{2\sigma_u^2}\varepsilon'[\frac{\iota_T\iota_T'}{T}\otimes[(B_N'B_N)^{-1} + (T\eta)I_N]^{-1}]\varepsilon$$
$$-\frac{1}{2\sigma_u^2}\varepsilon'[(I_T - \frac{\iota_T\iota_T'}{T})\otimes(B_N'B_N)]\varepsilon\ ,$$

with $\varepsilon = y - X\beta$.

A second special case is the the spatial SUR model with spatial SAR error autocorrelation. Its error variance covariance matrix is given by (19.22). The required determinant and inverse for the log-likelihood are (see Anselin, 1988a, p. 143):

$$|B_{NT}^{-1}(\Sigma_T\otimes I_N)B_{NT}^{-1'}| = |\Sigma_T|^N|B_{NT}|^{-2}\ ,$$

and,

$$[B_{NT}^{-1}(\Sigma_T\otimes I_N)B_{NT}^{-1'}]^{-1} = B_{NT}'[\Sigma_T^{-1}\otimes I_N]B_{NT}\ .$$

Furthermore, due to the block-diagonal structure of B_{NT}:

$$\ln|B_{NT}| = \sum_t \ln|I_N - \theta_t W_N|\ .$$

The log-likelihood for this model then follows as:

$$L = -\frac{N}{2}\ln|\Sigma_T| + \sum_t \ln|I_N - \theta_t W_N|$$
$$-\frac{1}{2}\varepsilon' B_{NT}'(\Sigma_T^{-1}\otimes I_N)B_{NT}\varepsilon\ ,$$

with $B_{NT}\varepsilon$ corresponding to the residuals from spatially filtered dependent and explanatory variables, $[I_{NT} - (\Theta_T\otimes W_N)](y - X\beta)$, a generalization of the pooled model case.

19.4.2 Instrumental Variables and GMM

As an alternative to reliance on an often unrealistic assumption of normality and to avoid some of the computational problems associated with the Jacobian term in ML estimation, instrumental variables and GMM methods have been suggested for single cross-section spatial regression models (e.g., Anselin, 1988a, 1990; Kelejian and Robinson, 1993; Kelejian and Prucha, 1998, 1999; Conley, 1999). These can be extended to the panel data setting. We will consider the spatial lag and spatial error models in turn.

19.4.2.1 Spatial Lag Models

The endogeneity of the spatially lagged dependent variable suggests a straightforward instrumental variables strategy in which the spatially lagged (exogenous) explanatory variables WX are used as instruments (Kelejian and Robinson, 1993; Kelejian and Prucha, 1998; and also Lee, 2003 for the choice of optimal instruments). This applies directly to the spatial lag in the pooled model, where the instruments would be $(I_T \otimes W_N)X$ (with X as a stacked $NT \times (K-1)$ matrix, excluding the constant term).

A special case is the spatial SUR model with a spatial lag term, (19.19). Following the same approach as taken in the single cross-section, consider the spatially lagged dependent variable and the explanatory variables in each equation grouped into a matrix $Z_t = [W_N y_t \; X_t]$, with parameter vector $\gamma_t = [\rho_t \; \beta_t']'$. The individual Z_t terms can be stacked into a $NT \times T(K+1)$ matrix Z, using the same setup as in (19.18), with a matching stacked coefficient vector γ. For each equation, construct a matrix of instruments, $H_t = [X_t \;\; W_N X_t]$, stacked in block-diagonal form into H. With a consistent estimate for the error variance covariance matrix, $\hat{\Sigma}_T \otimes I_N$, the model parameters can be estimated by means of the IV estimator with a general non-spherical error variance (Anselin, 1988a, p. 146):

$$\hat{\gamma} = \left[Z'H[H'(\hat{\Sigma}_T \otimes I_N)H]^{-1}H'Z\right]^{-1} Z'H[H'(\hat{\Sigma}_T \otimes I_N)H]^{-1}H'y \tag{19.38}$$

with an estimate for the coefficient variance as:

$$\text{Var}[\hat{\gamma}] = \left[Z'H[H'(\hat{\Sigma}_T \otimes I_N)H]^{-1}H'Z\right]^{-1} .$$

This suggests an iterative spatial three stages least squares estimator (S3SLS): first estimate each regression using spatial 2SLS (S2SLS); use the S2SLS residuals to obtain a consistent estimate of $\hat{\Sigma}$; and finally use $\hat{\Sigma}$ in (19.38). Consistency and asymptotic normality of the spatial generalized IV estimator can be based on the arguments developed for the cross-sectional S2SLS case (Kelejian and Robinson, 1993; Kelejian and Prucha, 1998).

19.4.2.2 Spatial Error Models

The spatially weighted least squares result (19.37) for the regression parameters in the pooled model with SAR errors also holds in a more general setting, without assuming normality. As long as a consistent estimator for the *nuisance* parameter θ can be obtained, the FGLS estimator will also be consistent for β.

In the single cross-section, a consistent estimator can be constructed from a set of moment conditions on the error terms, as demonstrated in the Kelejian–Prucha generalized moments (KPGM) estimator (Kelejian and Prucha, 1999). These conditions can be readily extended to the pooled model, by replacing the single equation spatial weights by their pooled counterparts $(I_T \otimes W_N)$. The point of departure is the stacked vector of SAR errors:

$$\varepsilon = \theta(I_T \otimes W_N)\varepsilon + u\,,$$

where both ε and u are $NT \times 1$ vectors, and $u \backsim \text{IID}[0, \sigma_u^2 I_{NT}]$.

The three KPGM moment conditions (Kelejian and Prucha, 1999, p. 514) pertain to the idiosyncratic error vector u. Extending them to the pooled setting yields:

$$\begin{aligned} \text{E}[\frac{1}{NT}u'u] &= \sigma_u^2 \\ \text{E}[\frac{1}{NT}u'(I_T \otimes W_N')(I_T \otimes W)u] &= \frac{1}{N}\sigma_u^2 \text{tr}(W_N'W_N) \\ \text{E}[\frac{1}{NT}u'(I_T \otimes W_N)u] &= 0\,, \end{aligned}$$

where tr is the matrix trace operator and use is made of $\text{tr}(I_T \otimes W_N'W_N) = T\text{tr}W_N'W_N$, and $\text{tr}(I_T \otimes W_N) = 0$.

The estimator is made operational by substituting $u = \varepsilon - \theta(I_t \otimes W_N)\varepsilon$, and replacing ε by the regression residuals. The result is a system of three equations in θ, θ^2 and σ_u^2, which can be solved by nonlinear least squares (for technical details, see Kelejian and Prucha, 1999). Under some fairly general regularity conditions, substituting the consistent estimator for θ into the spatial FGLS (19.37) will yield a consistent estimator for β. Recently, this approach has been extended to the error components model with spatial error dependence (19.29), yielding a system of six moment equations (for details, see Kapoor et al., 2007).

19.5 Testing for Spatial Dependence

Testing for spatial effects in spatial panel models centers on the null hypotheses $\text{H}_0 : \rho = 0$ and/or $\text{H}_0 : \theta = 0$ in the various models that include spatial lag terms or spatial error autocorrelation. Arguably, the preferred approach is based on Lagrange Multiplier (LM) or Rao Score (RS) tests, since these only require estimation of the model under the null, avoiding the complexities associated with ML estimation (for a recent review, see Anselin, 2001a). The test statistics developed for the single cross-section case can be readily extended to the pooled model. In addition, specialized diagnostics have been developed to test for spatial effects in spatial SUR (Anselin, 1988b), and for error components models (Anselin, 1988a; Baltagi et al., 2003, 2006, 2007). More recently, a strategy has been suggested to test for general unspecified cross-sectional dependence (Pesaran, 2004).

We focus our attention on the LM tests and first briefly review the generic case. This is followed by an illustration of applications of the LM principle to tests against error correlation in the spatial SUR and error components models.

19.5.1 Lagrange Multiplier Tests for Spatial Lag and Spatial Error Dependence in Pooled Models

The results for the pooled models follow as straightforward generalizations of the single cross-section case, with proper adjustments for the spatial weights matrix and weights matrix traces. Consider the pooled regression model (19.2) as the point of departure, with $e = y - X\hat{\beta}$ as a $NT \times 1$ vector of regression residuals.

The single cross-section Lagrange Multiplier test statistic for spatial error correlation, LM_E (Burridge, 1980), which is asymptotically distributed as $\chi^2(1)$, is readily extended to the pooled model with spatial weights matrix $(I_T \otimes W_N)$ as:

$$LM_E = \frac{[e'(I_T \otimes W_N)e/(e'e/NT)]^2}{\text{tr}[(I_T \otimes W_N^2) + (I_T \otimes W_N'W_N)]}$$

or, using simplified trace terms:

$$LM_E = \frac{[e'(I_T \otimes W_N)e/(e'e/NT)]^2}{T\,\text{tr}(W_N^2 + W_N'W_N)} .$$

Similarly, the single cross-section LM test statistic for a spatial lag alternative, LM_L (Anselin, 1988a), becomes:

$$LM_L = \frac{[e'(I_T \otimes W_N)y/(e'e/NT)]^2}{[(W\hat{y})'M(W\hat{y})/\hat{\sigma}^2] + T\,\text{tr}(W_N^2 + W_N'W_N)}$$

with $W\hat{y} = (I_T \otimes W_N)X\hat{\beta}$ as the spatially lagged predicted values in the regression, and $M = I_{NT} - X(X'X)^{-1}X'$. This statistic is also asymptotically distributed as $\chi^2(1)$.

This simple approach can be generalized to account for more realistic error variance structures, such as heteroskedasticity across the time periods, in the same manner that heteroskedasticity is included in test statistics for the single cross-section (see, e.g., Kelejian and Robinson, 1998). Alternatively, each of the test statistics can be robustified against the alternative of the other form, using the standard approach (see Anselin et al., 1996).

19.5.2 Testing for Spatial Error Correlation in Panel Data Models

19.5.2.1 Spatial SUR Model

In the spatial SUR model (19.15–19.17), the LM test statistics are based on the residuals from a standard ML or FGLS estimation. In contrast to the pooled model, the null hypothesis pertains to T parameter constraints, $H_0 : \theta_1 = \ldots = \theta_T = 0$ for the spatial error alternative.

To construct the statistic, consider a $N \times T$ matrix E with the $N \times 1$ individual equation residual vectors as columns. The LM_E test statistic then follows as (Anselin, 1988b):

$$LM_E = \iota_T'(\hat{\Sigma}_T^{-1} * E'W_N E)J^{-1}(\hat{\Sigma}_T^{-1} * E'W_N E)'\iota_T$$

with $*$ as the Hadamard product, and

$$J = [\mathrm{tr}(W_N^2)]I_T + [\mathrm{tr}(W_N'W_N)](\hat{\Sigma}_T^{-1} * \hat{\Sigma}_T)$$

The LM_E statistic is distributed asymptotically as $\chi^2(T)$.

19.5.2.2 Error Components Models

In the error components model with spatial autoregressive errors (19.25–19.26), the null hypothesis is $\mathrm{H}_0 : \theta = 0$. A LM test statistic can be constructed from the residuals obtained by estimating the standard error components model by FGLS or ML. With e as the $NT \times 1$ residual vector, and, to simplify notation, with $\hat{\kappa} = (\hat{\sigma}_\mu^2/\hat{\sigma}_u^2)/[1 + T(\hat{\sigma}_\mu^2/\hat{\sigma}_u^2)]$, the test statistic follows as (Anselin, 1988a, p. 155):

$$LM_E = \frac{\left[(1/\hat{\sigma}_u^2)e'[[I_T + \hat{\kappa}(T\hat{\kappa} - 2)\iota_T\iota_T'] \otimes W_N]e\right]^2}{p},$$

with $p = (T^2\hat{\kappa}^2 - 2\hat{\kappa} + T)(\mathrm{tr}W_N^2 + \mathrm{tr}W_N'W_N)$. It is distributed asymptotically as $\chi^2(1)$.

When the point of departure is not the error components model, but the pooled specification (19.2), both the error component and the spatial parameter can be considered as part of the null hypothesis, and a number of interesting combinations result. The resulting tests can be classified as marginal, joint or conditional, depending on which combinations of parameters restrictions are considered (Baltagi et al., 2003).

Specifically, marginal tests would be on either $\mathrm{H}_0 : \theta = 0$ (the spatial parameter) or on $\mathrm{H}_0 : \sigma_\mu^2 = 0$ (the error component), based on the residuals of the pooled model. A joint test is on $\mathrm{H}_0 : \theta = \sigma_\mu^2 = 0$, and conditional tests are for $\mathrm{H}_0 : \theta = 0$ (assuming $\sigma_\mu^2 \geq 0$), or $\mathrm{H}_0 : \sigma_\mu^2 = 0$ (assuming θ may or may not be zero). Each case yields a LM statistic using the standard principles applied to the proper likelihood function (for details, see Baltagi et al., 2003). This rationale can be further extended to include a time-wise dependent process with parameter ϕ, as in (19.32–19.33) (for detailed derivations, see Baltagi et al., 2007).

19.6 Conclusions

The econometrics of panel data models with spatial effects constitutes an active area of research, as evidenced by a growing number of recent papers on the topic. The focus to date has been primarily on theoretical and methodological aspects. Arguably, the dissemination of these methods to empirical practice has been hampered

by the lack of ready to use software. None of the standard econometric packages include built-in facilities to carry out single cross-section spatial econometrics, let alone spatial panel econometrics.

For single cross-section spatial econometrics, there are now several software resources available, ranging from freestanding packages such as GeoDa (Anselin et al., 2006), to collections of routines in Matlab (James LeSage's collection of routines at http://www.spatialeconometrics.com) and R (Bivand, 2002). However, apart from a few Matlab routines for spatial fixed effects models developed by Paul Elhorst (see http://www.spatialeconometrics.com), the situation is rather bleak for panel spatial econometrics in general. A promising development in this regard is the effort under the auspices of the U.S. Center for Spatially Integrated Social Science (CSISS) to develop software for spatial econometrics in the open source Python language. The PySpace collection of modules that is currently under active development includes the basic tests and estimation methods for the pooled panel model as well as the spatial SUR model (Anselin and Le Gallo, 2004).

While much progress has been made, many areas remain where very little insight has been gained into the complexities that result from explicitly introduction spatial dependence and spatial heterogeneity into panel data models. Directions with particular promise for future research would be the extension to models with discrete dependent variables. Also of particular interest to applied researchers would be greater insight into the trade offs involved in using strategies for general cross-sectional dependence relative to the use of parameterized spatial processes.

It is hoped that the review provided in the current chapter may provide a stimulus and resource to theoretical and applied researchers alike to aid in pursuing these directions in the future.

Acknowledgments The research by Luc Anselin and Julie Le Gallo was supported in part by the U.S. National Science Foundation Grant BCS-9978058 to the Center for Spatially Integrated Social Science (CSISS).

References

Anselin, L. (1988a). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Anselin, L. (1988b). A test for spatial autocorrelation in seemingly unrelated regressions. *Economics Letters*, 28:335–341.

Anselin, L. (2001a). Rao's score test in spatial econometrics. *Journal of Statistical Planning and Inference*, 97:113–139.

Anselin, L. (2001b). Spatial econometrics. In Baltagi, Badi, editor, *A Companion to Theoretical Econometrics*, pages 310–330. Blackwell, Oxford.

Anselin, L. (2002). Under the hood. Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 27(3):247–267.

Anselin, L. (2003). Spatial externalities, spatial multipliers and spatial econometrics. *International Regional Science Review*, 26(2):153–166.

Anselin, L. and Bera, A. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In Ullah, Amman and Giles, David E.A., editors, *Handbook of Applied Economic Statistics*, pages 237–289. Marcel Dekker, New York.

Anselin, L., Bera, A., Florax, Raymond J.G.M., and Yoon, M. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, 26:77–104.

Anselin, L. and Florax, Raymond J.G.M. (1995). *New Directions in Spatial Econometrics*. Springer-Verlag, Berlin.

Anselin, L., Florax, Raymond J.G.M., and Rey, Sergio J. (2004). Econometrics for spatial models, recent advances. In Anselin, Luc, Florax, Raymond J.G.M., and Rey, Sergio J., editors, *Advances in Spatial Econometrics. Methodology, Tools and Applications*, pages 1–25. Springer-Verlag, Berlin.

Anselin, L. and Le Gallo, J. (2004). *Panel Data Spatial Econometrics with PySpace*. Spatial Analysis Laboratory (SAL). Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL.

Anselin, L. and Moreno, R. (2003). Properties of tests for spatial error components. *Regional Science and Urban Economics*, 33(5):595–618.

Anselin, L., Syabri, I., and Kho, Y. (2006). Geoda, an introduction to spatial data analysis. *Geographical Analysis*. 38(1):5–22.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press, Oxford, United Kingdom.

Baltagi, Badi H. (2001). *Econometric Analysis of Panel Data (Second Edition)*. John Wiley & Sons, Chichester, United Kingdom.

Baltagi, Badi H., Egger, P., and Pfaffermayr, M. (2006). A generalized spatial panel data model with random effects. Working paper, Syracuse University, Syracuse, NY.

Baltagi, Badi H., Song, Seuck H., Jung, Byoung C., and Koh, W. (2007). Testing for serial correlation, spatial autocorrelation and random effects using panel data. *Journal of Econometrics,* 140(1):5–51.

Baltagi, Badi H., Song, Seuck H., and Koh, W. (2003). Testing panel data regression models with spatial error correlation. *Journal of Econometrics*, 117:123–150.

Banerjee, S., Carlin, Bradley P., and Gelfand, Alan E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, FL.

Barry, Ronald P. and Pace, R. Kelley (1999). Monte Carlo estimates of the log determinant of large sparse matrices. *Linear Algebra and its Applications*, 289:41–54.

Bivand, R. (2002). Spatial econometrics functions in R: Classes and methods. *Journal of Geographical Systems*, 4:405–421.

Brock, William A. and Durlauf, Steven N. (2001). Discrete choice with social interactions. *Review of Economic Studies*, 68(2):235–260.

Brueckner, Jan K. (2003). Strategic interaction among governments: An overview of empirical studies. *International Regional Science Review*, 26(2):175–188.

Burridge, P. (1980). On the Cliff-Ord test for spatial autocorrelation. *Journal of the Royal Statistical Society B*, 42:107–108.

Case, Anne C. (1991). Spatial patterns in household demand. *Econometrica*, 59:953–965.

Case, Anne C. (1992). Neighborhood influence and technological change. *Regional Science and Urban Economics*, 22:491–508.

Case, Anne C., Rosen, Harvey S., and Hines, James R. (1993). Budget spillovers and fiscal policy interdependence: Evidence from the states. *Journal of Public Economics*, 52:285–307.

Casetti, E. (1997). The expansion method, mathematical modeling, and spatial econometrics. *International Regional Science Review*, 20:9–33.

Chen, X. and Conley, Timothy G. (2001). A new semiparametric spatial model for panel time series. *Journal of Econometrics*, 105:59–83.

Cliff, A. and Ord, J. Keith (1981). *Spatial Processes: Models and Applications*. Pion, London.

Coakley, J., Fuentes, A.-M., and Smith, R. (2002). A principal components approach to cross-section dependence in panels. Working Paper, Department of Economics, Birkbeck College, University of London, London, United Kingdom.

Conley, Timothy G. (1999). GMM estimation with cross-sectional dependence. *Journal of Econometrics*, 92:1–45.

Conley, Timothy G. and Ligon, E. (2002). Economic distance, spillovers and cross country comparisons. *Journal of Economic Growth*, 7:157–187.

Conley, Timothy G. and Topa, G. (2002). Socio-economic distance and spatial patterns in unemployment. *Journal of Applied Econometrics*, 17:303–327.

Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94:1330–1340.

Cressie, N. (1993). *Statistics for Spatial Data.* Wiley, New York.

Driscoll, John C. and Kraay, Aart C. (1998). Consistent covariance matrix estimation with spatially dependent panel data. *The Review of Economics and Statistics*, 80:549–560.

Druska, V. and Horrace, William C. (2004). Generalized moments estimation for spatial panel data: Indonesian rice farming. *American Journal of Agricultural Economics*, 86(1):185–198.

Dubin, R. (1988). Estimation of regression coefficients in the presence of spatially autocorrelated errors. *Review of Economics and Statistics*, 70:466–474.

Dubin, R. (1995). Estimating logit models with spatial dependence. In Anselin, Luc and Florax, Raymond J.G.M., editors, *New Directions in Spatial Econometrics*, pages 229–242. Springer-Verlag, Berlin.

Elhorst, J. Paul (2001). Dynamic models in space and time. *Geographical Analysis*, 33:119–140.

Elhorst, J. Paul (2003). Specification and estimation of spatial panel data models. *International Regional Science Review*, 26(3):244–268.

Fazekas, I., Florax, R., and Folmer, H. (1994). On maximum likelihood estimators of parameters of spatio-temporal econometric models. Technical Report No. 109/1994, Kossuth University, Debrecen, Hungary.

Florax, Raymond J.G.M. and Van Der Vlist, Arno J. (2003). Spatial econometric data analysis: Moving beyond traditional models. *International Regional Science Review*, 26(3):223–243.

Fotheringham, A. Stewart, Brunsdon, C., and Charlton, M. (2002). *Geographically Weighted Regression*. John Wiley, Chichester.

Gamerman, D., Moreira, Ajax R.B., and Rue, H. (2003). Space-varying regression models: Specifications and simulation. *Computational Statistics & Data Analysis*, 42(3):513–533.

Gelfand, Alan E., Kim, H.-J., Sirmans, C.F., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98:387–396.

Giacomini, R. and Granger, Clive W.J. (2004). Aggregation of space-time processes. *Journal of Econometrics*, 118:7–26.

Glaeser, Edward L., Sacerdote, Bruce I., and Scheinkman, Jose A. (2002). The social multiplier. Technical Report 9153, NBER, Cambridge, MA 02138.

Haining, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge.

Hsiao, C. (1986). *Analysis of Panel Data.* Cambridge University Press, Cambridge.

Hsiao, C. and Pesaran, M. Hashem (2008). Random coefficient panel data models. In Matyas L. and Sevestre P., editors, *The Econometrics of Panel Data. Kuwer Academic Publishers*, Dordrecht.

Hsiao, C., Pesaran, M. Hashem, and Tahmiscioglu, A. Kamil (2002). Maximum likelihood estimation of fixed effects dynamic panel models covering short time periods. *Journal of Econometrics*, 109:107–150.

Kapoor, M., Kelejian, Harry H., and Prucha, Ingmar R. (2007). Panel data models with spatially correlated error components. *Journal of Econometrics*, 140(1):97–130.

Kelejian, Harry H. and Prucha, I. (1998). A generalized spatial two stage least squares procedures for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics*, 17:99–121.

Kelejian, Harry H. and Prucha, I. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, 40:509–533.

Kelejian, Harry H. and Robinson, Dennis P. (1993). A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model. *Papers in Regional Science*, 72:297–312.

Kelejian, Harry H. and Robinson, Dennis P. (1995). Spatial correlation: A suggested alternative to the autoregressive model. In Anselin, Luc and Florax, Raymond J.G.M., editors, *New Directions in Spatial Econometrics*, pages 75–95. Springer-Verlag, Berlin.

Kelejian, Harry H. and Robinson, Dennis P. (1998). A suggested test for spatial autocorrelation and/or heteroskedasticity and corresponding Monte Carlo results. *Regional Science and Urban Economics*, 28:389–417.

Lee, L.-F. (2002). Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. *Econometric Theory*, 18(2):252–277.

Lee, L.-F. (2003). Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. *Econometric Reviews*, 22:307–335.

Magnus, J. (1978). Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance covariance matrix. *Journal of Econometrics*, 7:281–312. Corrigenda, *Journal of Econometrics* 10, 261.

Manski, Charles F. (1993). Identification of endogenous social effects: The reflexion problem. *Review of Economic Studies*, 60:531–542.

Manski, Charles F. (2000). Economic analysis of social interactions. *Journal of Economic Perspectives*, 14(3):115–136.

Mardia, K.V. and Goodall, C. (1993). Spatio-temporal analyses of multivariate environmental monitoring data. In Patil, G.P. and Rao, C.R., editors, *Multivariate Environmental Statistics*, pages 347–386. Elsevier, Amsterdam.

Mardia, K.V. and Marshall, R.J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71:135–146.

Ord, J. Keith (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70:120–126.

Pace, R. Kelley and Barry, R. (1997). Sparse spatial autoregressions. *Statistics and Probability Letters*, 33:291–297.

Paelinck, J. and Klaassen, L. (1979). *Spatial Econometrics*. Saxon House, Farnborough.

Pesaran, M. Hashem (2002). Estimation and inference in large heterogenous panels with cross section dependence. DAE Working Paper 0305 and CESifo Working Paper no. 869, University of Cambridge, Cambridge, United Kingdom.

Pesaran, M. Hashem (2004). General diagnostic tests for cross section dependence in panels. Working paper, University of Cambridge, Cambridge, United Kingdom.

Rey, Sergio J. and Montouri, Brett D. (1999). US regional income convergence: A spatial econometrics perspective. *Regional Studies*, 33:143–156.

Smirnov, O. and Anselin, L. (2001). Fast maximum likelihood estimation of very large spatial autoregressive models: A characteristic polynomial approach. *Computational Statistics and Data Analysis*, 35:301–319.

Stein, Michael L. (1999). *Interpolation of Spatial Data, Some Theory for Kriging*. Springer-Verlag, New York.

Topa, G. (2001). Social interactions, local spillover and unemployment. *Review of Economic Studies*, 68(2):261–295.

Upton, Graham J. and Fingleton, B. (1985). *Spatial Data Analysis by Example. Vol. 1: Point Pattern and Quantitative Data*. Wiley, New York.

Waller, L., Carlin, B., and Xia, H. (1997a). Structuring correlation within hierarchical spatio-temporal models for disease rates. In Grégoire, T., Brillinger, D., Russek-Cohen, P., Warren, W., and Wolfinger, R., editors, *Modeling Longitudinal and Spatially Correlated Data*, pages 309–319. Springer-Verlag, New York.

Waller, L., Carlin, B., Xia, H., and Gelfand, A. (1997b). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92:607–617.

Wikle, Christopher K., Berliner, L. Mark, and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, 5:117–154.

Part III
Applications

Chapter 20
Foreign Direct Investment: Lessons from Panel Data

Pierre Blanchard, Carl Gaigné and Claude Mathieu

20.1 Introduction

Since the 1980s, foreign direct investment (FDI) flows have grown substantially, especially throughout OECD countries (UNCTAD (2002)). The average share of FDI outflows in GDP went from around 2% in 1985 to almost 11% by the end of the 1990s. In 2000, OECD countries were the source of 90% FDI flows, and the recipient of 79%. International corporations are now major actors in international trade since their contribution to global production climbed to 11% in 2001. It is not surprising that, over the last two decades, FDI has spawned a significant amount of academic research and the literature continues to grow at an impressive rate. The empirical literature has expanded at a rapid pace in many different directions. Regardless of the question studied, the nature of the problem itself generally requires using panel data estimation methods because flows (or stocks) of FDI between pairs of countries (or between country-industry pairs) are analyzed for one or several time period. The purpose of this chapter is to provide a selective survey of the empirical literature using panel data.

As we will see in the next section, the theoretical literature has identified two dimensions acting upon the structure of FDI. The main components of firm/industry characteristics are transport costs, plant scale economies and factor intensities whereas market size, tariff levels and factor abundance are the main components of country features. As a result, using the econometrics of panel data is a natural way to evaluate the determinants to FDI. This issue is discussed in Sect. 20.3. In Sect. 20.4,

Pierre Blanchard
Erudite, Faculté de Sciences Economiques et de Gestion, Université Paris XII Val de Marne, 61 Av. du Général de Gaulle, 94010 Créteil Cédex, France, e-mail: blanchard@univ-paris12.fr

Carl Gaigné
INRA, UMR1302, SMART, F-35000 Rennes, France, e-mail: gaigne@rennes.inra.fr

Claude Mathieu
Erudite, Faculté de Sciences Economiques et de Gestion, Université Paris XII Val de Marne, 61 Av. du Général de Gaulle, 94010 Créteil Cédex, France, e-mail: mathieu@univ-paris12.fr

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

we present three types of empirical studies using the econometrics of panel data. The first type concerns the trade-off between producing at home or abroad. The second type studies more precisely the role of trade policies (anti-dumping, threat of protectionism, custom union) in the decision to establish an additional plant in a foreign country. The last type focuses on the impact of financial factors on the level of FDI. In the last section, we discuss recent econometric issues related to estimating FDI models using panel data.

Before presenting the micro-foundations of the decision to produce abroad, we have to define foreign direct investment. FDI refers to investments by multinational firms (MNF) in affiliates or subsidiaries. It consists of two broad categories: (i) direct net transfers from the parent company to the foreign affiliate, either through equity or debt; and (ii) reinvested earnings by a foreign affiliate. FDI is generally thought as a real capital flow between countries, the main interest in our analysis. Still, statistical information on FDI involves financial flows that do not necessarily correspond to an international allocation of productive capital. Indeed, FDI is comprised of several types of capital. First, it contains real investment in plants and equipment, either in the form of new plants and equipment or plant expansion. Second, a major part of FDI consists of the financial flows associated with mergers and acquisitions. This implies an ownership change in the absence of any real investment. OECD (2000) estimates suggest that mergers and acquisitions account for more than 60% of all FDI in developed countries. Others components of FDI are joint ventures and equity increases. The latter component typically comprises investment in financial capital. The distinction between the various types of FDI is important because the different components may have different explanations.

20.2 A Simple Model of FDI

In the 1980s, trade economists proposed refinements of the factor-proportions approach to explain the emergence of multinational corporations (e.g. Helpman (1984)). They determine the conditions under which firms have an incentive to become a "vertical" multinational, that is to separate headquarters from plant. A vertical multinational activity arises between countries that differ significantly in relative endowments. However, in order to explain the existence of foreign direct investments among similar countries, an alternative approach has been proposed by different authors (e.g. Markusen (1984)). The purpose is to determine the conditions under which firms produce the same product in multiple plants, serving local markets by local production. A firm will probably be a horizontal multinational when trade costs are relatively high and plant-level scale economies are low enough. This theoretical literature on FDI is generally characterized by general equilibrium models (see Markusen (1995)). However, in order to make the results accessible, we do not develop a general equilibrium model of plant location in this section. The objective of this section is to show basic mechanisms at work by developing a simple model of foreign investments, which is close to Markusen (2002). We will see

how the main characteristics of technologies and countries interact to determine the choice of firms to engage in FDI and the type of FDI (horizontal or vertical). Technology features include plant-level and firm-level scale economies whereas country features include trade costs and global market size as well as differences in market size and marginal costs of production.

20.2.1 Assumptions and Preliminary Results

Consider one good produced by a single firm and sold in two markets/countries (h and f). Countries may differ in population size and/or in technology. The production of the good implies two types of sunk cost: a plant-specific fixed cost (G) by production unit and a firm-specific cost (F). Consumers are internationally immobile and both markets are segmented. The firm practices third degree price discrimination without threat of arbitrage by consumers. There are three alternatives modes of serving both markets. (i) By a national firm with a single plant located in country h (*type-n*). The national firm serves country f by exporting, which implies operational costs t such as transportation costs as well as other nontariff trade barriers. We assume that t is symmetric between countries. (ii) By a horizontal multinational with two plants located in both countries (*type-h*). The horizontal multinational serves country f by establishing a subsidiary abroad, which implies further plant-specific fixed cost G. (iii) By a vertical multinational with the headquarter located in country h and one plant in country f that serves both markets (*type-v*).

The inverse demand function in each country is given by

$$p_{ij} = a - (b/L_i)q_{ij} \tag{20.1}$$

where p_{ij}, q_{ij} are price and quantity of the good produced in country $i = h, f$ and sold in country $j = h, f$. In addition L_i is the population in country $i = h, f$. We assume that $a, b > 0$.

The expression of profits of a *type-n* firm is expressed as follows:

$$\pi^n = (a-(b/L_h)q_{hh})q_{hh} + [(a-(b/L_f)q_{hf}) - t]q_{hf} - c_h(q_{hh}+q_{hf}) - G - F \tag{20.2}$$

where c_h is the marginal cost of production prevailing in country h, and F a firm-specific fixed cost. By solving the first-order conditions, the profit-maximizing output in both markets is given by,

$$q^*_{hh} = \frac{a-c_h}{2b}L_h \quad \text{and} \quad q^*_{hf} = \frac{a-c_h-t}{2b}L_h \tag{20.3}$$

Consider now that the firm is a horizontal multinational. Its profit function is

$$\pi^h = [(a-(b/L_h)q_{hh}) - c_h]q_{hh} - G + [(a-(b/L_f)bq_{ff}) - c_f]q_{ff} - G - F \tag{20.4}$$

The supply on the foreign market corresponds to

$$q_{ff}^{*} = \frac{a - c_f}{2b} L_f \tag{20.5}$$

Note that the supply on the domestic market is q_{hh}^{*} from (20.3).

Finally, when the multinational adopts a *type-v* structure, its profit equation is expressed as follows:

$$\pi^{v} = [(a - (b/L_h)q_{fh}) - c_f - t]q_{fh} + [(a - (b/L_f)q_{ff}) - c_f]q_{ff} - G - F \tag{20.6}$$

Maximizing (20.6) gives the export sales from country f to country h:

$$q_{fh}^{*} = \frac{a - c_f - t}{2b} L_h \tag{20.7}$$

whereas the sales in country f (q_{ff}^{*}) are given by (20.5).

We can now summarize the total profits under the three alternative modes of serving country h and f by introducing (20.3), (20.5) and (20.7) in (20.2), (20.4) and (20.6), respectively,

$$\pi^{n} = \left(\frac{a - c_h}{2b}\right)^2 L_h + \left(\frac{a - c_h - t}{2b}\right)^2 L_f - G - F \tag{20.8}$$

$$\pi^{h} = \left(\frac{a - c_h}{2b}\right)^2 L_h + \left(\frac{a - c_f}{2b}\right)^2 L_f - 2G - F \tag{20.9}$$

$$\pi^{v} = \left(\frac{a - c_f - t}{2b}\right)^2 L_h + \left(\frac{\alpha - c_f}{2b}\right)^2 L_f - G - F \tag{20.10}$$

20.2.2 Technology and Country Characteristics as Determinants of FDI

The previous three profit equations enable us to determine the main factors that determine the choice for a firm about whether or not to engage in foreign investment and the type of FDI (horizontal or vertical). To simplify the analysis, we assume that $a > c_h \geq c_f$ and $L_h \geq L_f$ where a is sufficiently large as well as L_f. We consider four configurations: (i) the characteristics of both countries are identical; (ii) the size of the home market is larger; (iii) the marginal cost of production is lower in the foreign country; (iv) combination of cases (ii) and (iii).

(i) First, we assume that countries are identical with respect to technology and factor endowments ($c_h = c_f = c$ and $L_h = L_f = L$). So, we have $\pi^{n} = \pi^{v}$. Trivial calculations show that the firm decides to produce in both countries ($\pi^{h} > \pi^{n}$) if and only if trade costs are high enough or equivalently when $t > t^{nh}$ where

$$t^{nh} \equiv (a-c)\left(1-\sqrt{1-\frac{4bG}{L_f(a-c)^2}}\right) > 0 \tag{20.11}$$

The threshold value t^{nh} increases when G declines and decreases when the population size of countries (L) grows. In other words, *the firm will be likely a horizontal multinational when trade costs are high relatively to plant scale economies and when the markets to serve are large enough*.

(ii) Assuming now that countries are only different in population size with $c_h = c_f = c$ and $L_h > L_f$. Therefore, country h has an advantage in market size. In this case, regardless of values of trade costs, profits when the firm adopts a *type-n* structure is always superior to profits when it chooses a *type-v* structure ($\pi^n > \pi^v$). The critical value of trade costs above which the national firm becomes a horizontal multinational is identical to t^{nh}, except that L_f is now lower than L_h. As a result, it appears that, when the market size is higher at home, *convergence in population size between countries prompts the firm to establish a second plant abroad*.

(iii) We now consider the case where countries are only different in production costs with $c_h > c_f$ and $L_h = L_f = L$. Stated differently, country f has an advantage in production costs. In this configuration, we have $\pi^n < \pi^v$ regardless of trade costs. When production costs differ among countries, the firm has a strong incentive to become vertical multinational. In addition, the multinational produces in both countries if and only if $t > t^{vh}$ where

$$t^{vh} = (a-c_f)\left(1-\sqrt{1-\frac{4G}{b^2L(a-c_h)^2}}\right) \tag{20.12}$$

It is readily confirmed that *horizontal direct investments are favored when marginal costs converge*.

(iv) Finally, with $c_h > c_f$ and $L_h > L_f$, we consider the case where the advantage in market size benefits country h while the advantage in production costs benefits country f. This configuration is more complex because we must rank three profit equations: π^v, π^n and π^h. Figure 20.1 shows graphically the profits of each regime against trade costs. It is straightforward to check that $\pi^v(t=0) > \pi^n(t=0)$ and that profits in both structures (*type-v* and *-n*) decline when trade costs increase. In addition, we have $\pi^v(t=0) > \pi^h$. As a result, the multinational is more likely to have a vertical structure when trade costs are very low. Further, as profits do not vary with respect to trade costs when the firm is characterized by a *type-h* structure (see the dashed lines in Fig. 20.1), a horizontal FDI is more likely to take place when trade costs are high enough. Finally, the firm becomes national when trade costs take intermediate values. Note that the *relative* position of the profit curves depends on the size of plant scale economies (G). More precisely, a fall in G increases the profits more when the multinational is located in both countries than when the firm produces in a single country.

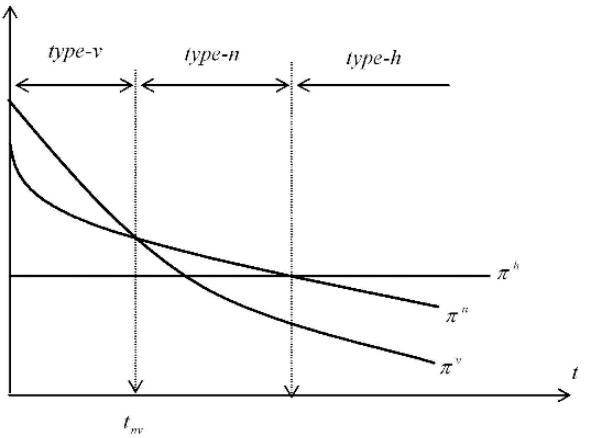

Fig. 20.1 Profit functions according to the firm type

Observe also that we have $\pi^v(t) > \pi^n(t)$ if and only if $t < t_{nv}$ where

$$t_{nv} \equiv (c_h - c_f)\frac{L_h + L_f}{L_h - L_f} \tag{20.13}$$

Then, the firm is likely to prefer to serve both countries from the foreign country when trade costs are low enough. In addition, when the size of markets diverges noticeably (L_h grows or L_f declines), the *type-v* firm is more likely to occur (t_{nv} increases). Consequently, the advantage in production costs dominates the advantage in market size when trade costs are sufficiently low, while the advantage in market size dominates the advantage in production costs when trade costs are high enough.

To summarize our analysis, we first recall the main conditions under which a firm engages in horizontal FDI: countries are similar in market size and in marginal production costs, the "world" demand is sufficiently high, the plant-specific fixed cost is low relative to the firm-specific fixed cost and trade costs are high enough. In addition, the firm is more likely to be a vertical multinational when trade costs are low enough and when the difference in production costs is sufficiently high. Finally, notice that FDI and trade are substitutes when multinationals are horizontal and complementary when multinationals are vertical.

20.3 Econometric Implementation and Data

The basic model of the previous section has allowed us to identify the factors acting upon the emergence of FDI at two levels: at firm/industry level (technology, plant scale economies, factor intensities and transport costs) and at country level (market size, tariff levels and factor endowments). As a result, panel data models have been

extensively used for analyzing the factors determining the international allocation of foreign investments.

20.3.1 A General Econometric Model

Ideally, in order to control for observed and unobserved heterogeneity between host and domestic countries and for time effects, we need to estimate triple indexed models (see Matyas (2001), Egger and Pfaffermayr (2003)). A basic specification[1] is, for instance,

$$\mathrm{FDI}_{ijt} = \mathbf{x}_{ijt}\beta + \alpha_i + \lambda_j + \gamma_t + \delta_{ij} + u_{ijt} \tag{20.14}$$

where FDI_{ijt} is the amount of outward FDI of country i (home) held in the country j (host) at year t and $\mathbf{x}_{ijt}$ is a vector of regressors.[2] As suggested by the theoretical model, the variables included in the regressors list may be: a measure of bilateral country size (e.g. the sum of bilateral GDP); an index of size similarity; one or more measure of differences in relative endowments (e.g. capital stock and/or skilled labor ratios between the home and host countries); a variable measuring trade costs (tariffs, distance, for example). Nevertheless, several variables are added in order to control for investment, political, financial risks, non tariff barriers, openness policy.... Because the specification takes into account the effect of "gravity" factors (e.g. market size, distance), this model is usually called the gravity model and is commonly used not only for FDI analysis but also for modelling trade between countries. The parameters α_i and λ_j are introduced in order to control for heterogeneity across (host and domestic) countries (due for instance to legal or cultural characteristics) whereas γ_t captures any time-specific effect common to all pairs of countries such as business cycle effects, or changes in the degree of openness across all economies. The term δ_{ij} accounts for all time-invariant effects between two given countries such as common language and common borders.[3] These effects are modelled either as fixed (fixed effects model) or as random (random effects model) or, in very few studies, with a random coefficients specification (e.g. Feinberg and Keane (2001) and Balestra and Negassi (1992)). The Hausman test is frequently used in order to choose between the fixed effects and the random effects specification.

A second reason for using panel data for estimating FDI models relies often on the necessity to take into account the correlation between contemporaneous FDI flows and those of the previous year due to adjustment and sunk costs. By

[1] A more general specification is given by Baltagi, Egger and Pfaffermayr (2003). See Sect. 20.4.

[2] Several variants of this specification are frequently used in applied works, e.g.: (i) one can explain bilateral FDI from country i to country j for a given year (FDI_{ij}). (ii) It is also possible to focus on FDI from a given home country to several host countries at time t (FDI_{jt}) or in sector s at time t for each host country j (FDI_{jst}); (iii) one can also model FDI from a parent firm i to affiliates j at time t (FDI_{ijt}).

[3] Note also that the country-pair effects may differ according to the direction of FDI (i.e. $\delta_{ij} \neq \delta_{ji}$) which can be tested for.

including FDI_{ijt-1} in the model, we have a dynamic specification of FDI. Although this allows us to distinguish between short-run and long-run effects, it creates a correlation between the lagged dependant variable and the error term. However in this case, the usual estimators (OLS, within, GLS) are biased and inconsistent (for short T). With panel data, this problem may be solved by transforming the model in first differences. Moreover, using the time dimension, we can quite easily find instruments for endogenous regressors, FDI_{ijt-1} of course, but also for other explanatory variables which may be endogenous in such a context, e.g. GDP, exchange rate.

Therefore, it is not surprising that a vast and recent econometric literature using panel data has emerged on these topics with a great variability in the estimation methods used.

20.3.2 FDI and Data Issues

Testing model (20.14) requires data that vary in different dimensions (firm/industry, country and time). There are two main types of data on foreign direct investment: (i) the balance of payments provides information on inward and outward flows of FDI and the stocks derived from accumulated FDI flows. Such data are available at country level and vary over time; (ii) the second type of data is about operations of individual multinational firm at home and abroad. We discuss the advantages and disadvantages of these two types of data sets.

Type (i). Different international institutions publish international data on FDI based on the balance of payments, such as the International Monetary Fund (IMF), the United Nations (via the UNCTAD World Investment Report) and the Organization for Economic Cooperation and Development (OECD). These data sets cover many countries but many of them deviate significantly from the international guidelines for the compilation of balance of payments and international investment position statistics in the IMF's Balance of Payment Manual (5th edition) and in the OECD's Benchmark Definition of Foreign Direct Investment (3rd edition). We choose to describe more precisely the data provided by the OECD (see Lipsey (2001), for a description of data from IMF and United Nations) for two main reasons: first, FDI between OECD countries represent more than 60% of the overall FDI; second, these countries are more in accordance with the recommendations of Survey of Implementation of International Methodological Standards for Direct Investment (SIMSDI) which is *a comprehensive study of data sources, collection methods, and dissemination and methodological practices for FDI statistics*. The Directorate for Financial, Fiscal and Enterprise Affairs of OECD yields statistics on FDI transactions and positions, published under the title International Direct Investment Statistics Yearbook. The flows and stocks of FDI are compiled by using the balance of payments and the international investment positions, respectively. Both data sets are available for inward and outward FDI by partner country and by

industry (according to ISIC Rev. 3 classifications) in the standard format defined by the international guidelines. A few OECD countries do not provide complete information and/or deviate from the agreed international standards established by the IMF. Moreover, the database covers 28 OECD countries over the 1980–2000 period. As a result, the cross-sectional comparability of the data is improving and balanced panels can be easily implemented (see IMF and OECD (2001) and Nicoletti, Golub, Hajkova, Mirza and Yoo (2003) for further details).

Type (ii). Although it is becoming more widely recognized that data need to account for heterogeneity between MNF, this type of information is still relatively scarce. Some countries collect information about inward FDI (France, Germany, Italy, among other) or outward FDI (Japan, see Falzoni (2000) for a description). With few exceptions, only the US and Sweden produce data for both outward and inward FDI.[4] Moreover, information about the characteristics relative to the parent companies and their affiliates are less frequent.[5] In fact, the US Bureau of Economic Analysis (BEA) provides the more extensive database about the operations of the affiliates and their parent companies. Indeed, available data give specific information about gross product, employment, wages and R&D expenditures for each domestic or foreign unit belonging to a MNF. In the same vein, the Research Institute of Industrial Economics (IUI) in Sweden compiles a dataset which is based on a questionnaire sent to all Swedish MNF, containing information on parent companies as well as on the operations of each individual subsidiary (see Braunerhjelm, Ekholm, Grundberg and Karpaty (1996) for a detailed description of this database). However, this survey is only implemented approximately every four years since the 1960s. This means it is difficult to build a panel with a time dimension. Even though the BEA only conducts benchmark surveys every 5 years, its annual surveys can be used to build a more detailed panel dataset. A more important limit is that this dataset covers only US bilateral activity. Note that this limit is applied to all MNF databases developed by the different national official statistical departments (see for example the database on intra-firm international trade of the Department for Industrial Studies and Statistics (SESSI) of the French Ministry for Economic Affairs, Finance and Industry) and also to data gathered by private agencies (see Head and Ries (2001) for the database of Toyo Keizai on Japanese MNF). In addition, studies testing the trade-off between US export and FDI use the ratio of US exports to the sales of US multinational affiliates as the dependent variable. Indeed, the BEA's database does not contain firms which are only exporters. Consequently, the multinational firm sales must be aggregated at the level of industries in order to make comparable FDI and export data.

[4] See Stephan and Pfaffmann (1998) and Lipsey (2001) for a description and a discussion of different national sources of FDI as in Sweden, Germany, Japan and Canada.

[5] Note that these characteristics are not necessarily comparable when they exist (see for example the survey about the foreign affiliates of the French Ministry of Finance in 2000).

20.4 Empirical Estimations: Selected Applications

20.4.1 Testing the Trade-Off Between FDI and Exports

20.4.1.1 FDI Versus Exports

A branch of the empirical literature on multinational production-location decisions has used the BEA database in order to study the determinants of FDI at the country/industry level. Within this literature, Brainard (1997) was the first to use direct industry- and country-specific measures of several determinants of FDI. Her objective is to test the determinants of the horizontal integration of multinationals (see Sect. 20.2). However, the author also controls for the possibility that multinational activity is motivated by gaining access to factor supplies (vertical integration). This work has been extended in two ways. First, Helpman, Melitz and Yeaple (2004) focus on the horizontal dimension of FDI location decision by taking into account the heterogeneity within sectors. In parallel, Yeaple (2003) controls more explicitly for the vertical dimension of FDI decisions by including the interaction between factor intensities and factor abundance. Combining the approaches of Brainard, Helpman et al. and Yeaple, the general empirical model is given by:

$$EXSH_{js} = \beta_0 + \sum_c \beta_c \mathbf{cft}_{js} + \sum_l \beta_l \mathbf{scale}_s + \beta_u UC_{js} + \nu_s + \lambda_j + \varepsilon_{js} \qquad (20.15)$$

where EXSH_{js} is, in the three papers, the ratio of total US sales of good s in country j to the sum of local affiliate sales and exports from US to that host country.[6] Note that Brainard and Yeaple consider also the share of US imports as a dependent variable. In this subsection, we only focus on outward foreign investment.

$\mathbf{cft}_{js}$ is a vector of trade costs, such as transport and insurance costs ($\mathrm{FREIGHT}_{js}$) and tariff barriers (TARIFF_{js}). In Brainard and Yeaple, $\mathrm{FREIGHT}_{js}$ is measured by the freight and insurance charges reported by importers to the US Bureau of Census to calculate freight factors such as the ratio of charges to import values.[7] The data on TARIFF_{js} comes from a 1988/1989 database by the General Agreement on Tariffs and Trade on *ad valorem* tariffs at the 3-digit SIC level of industry. In Helpman et al., $\mathrm{FREIGHT}_{js}$ is computed as the ratio of CIF imports into the US to FOB imports from the data presented by Feenstra (1997) whereas TARIFF_{js} is calculated at the BEA industry/country level.

$\mathbf{scale}_s$ stands for scale economies in each industry. Two types of scale economies must be distinguished: at the corporate level (CSCALE_s) and at the plant level (PSCALE_s). They correspond to G and F in our theoretical model, respectively. In Brainard, CSCALE_s is measured as the number of nonproductive workers in the

[6] Relatively to model (20.14), we deleted the subscript i since we have one home country (the USA) and introduced a further dimension (the sector) indexed by the subscript s. Hence, ν_s is a sector-specific component, capturing the sectoral characteristics of firms in each country that are unobservable or omitted from equation, but do not vary over country.

[7] No comparable data are available from exporters. The authors assume that transport costs are symmetric, which introduces measurement error in the outward estimates.

average US-based firm and, in Yeaple, as the average number of nonproduction employees at the firm level. In Brainard, $PSCALE_s$ is defined as the number of production employees in the median US plant ranked by value added, whereas Yeaple uses the average number of production workers in the US plants. Helpman et al. calculate the average number of non-production workers at the six-digit level. They then compute this measure for every three-digit level as the average of the within three-digit sectors, weighted by the six-digit level sales in this sector.

The variable UC_{js} stands for unit costs of production, introduced in order to control for factor-proportions differences. In Yeaple, this cost is a vector of variables that reflect a potential host country's unit cost of production by sector. Brainard considers variations only in country characteristics since the proxy used is the differential in per-worker income whereas Helpman et al. use only cross-industry variations in technology such as capital and R&D intensities.

As expected, the studies by Brainard and Helpman et al. suggest that the share of affiliate sales is increasing in trade barriers, transport costs and corporate sale economies and decreasing in production scale economies. The empirical analysis of Brainard suggests also that the comparative advantage motive for FDI is far less important. These findings support the horizontal model of FDI. However, by considering the interaction between factor abundance and factor intensities at highly disaggregated level, Yeaple shows that the comparative advantage in production cost is also a key determinant of FDI.

20.4.1.2 Horizontal Versus Vertical FDI

The previous empirical works suggest FDI decision may be motivated by both horizontal and vertical considerations. Recent theoretical works, which are called knowledge-capital models (henceforth KC model), show that vertical and horizontal firms can emerge simultaneously (cf. Markusen, Venables, Eby-Konan and Zhang (1996), Markusen (1997)). The main feature of the KC models is that exploitation of factor-price differences interacts with multi-plant scale economies to explain the decision and the nature of foreign investments. The results of these models, arising from simulations, relate the decision to produce abroad to country characteristics. Examples of empirical papers in this field are Carr, Markusen and Maskus (2001) and Blonigen, Davies and Head (2003) as well as Markusen and Maskus (2002a,b). Again, the data used in these empirical studies comes from the US BEA, although data are aggregated across industries to the country-level to form a panel of cross-country observations over the period 1986–1994.

The KC model is a more elaborate version of the model developed in Sect. 20.2. The central idea is that the services of knowledge and knowledge-generating activities can be spatially separated from production and supplied to production facilities. As knowledge-based services are more skill intensive than production, the multinational corporations have an incentive to locate the first (resp., second) activity in country where skilled (resp., unskilled) labor is relatively cheap. Consequently, the multinational firm can be vertically integrated. In addition, the output of

knowledge-base activities can be used simultaneously by multiple producers. Then, the existence of firm-level scale economies implies that the same products or services are produced in different countries. In this case, multinational firms can be horizontally integrated.

As in the theoretical model, three types of firms can emerge: (i) horizontal multinationals where plants are located in different countries and headquarters set up in the home country (*type-h*); (ii) vertical multinationals where production takes place in a single plant located in a foreign country while the headquarters are maintained in the home country (*type-v*); and (iii) national firms with a single plant where the production and knowledge-based services are located only in the home country (*type-n*). Given the new assumptions, the last regime is dominant in the country with the large market size and skilled worker endowment and when foreign investment barriers are high. *Type-h* firms are likely to be dominant if transport costs are high enough and if the nations are similar in relative factor endowments as well as in size. In other words, if countries are dissimilar either in size or in factor endowments, one nation will be favored. For example, if nations have the same factor endowments but differ in size, firms located in the larger country benefit from lower production cost. Thus, vertical multinationals may emerge when the home country is skilled-labor-abundant and small, unless trade costs from the host country (where production takes place) back to the parent country (where the headquarter is located) are too excessive.

Given mechanisms discussed above, Carr, Markusen and Maskus (2001) estimate the following equation:

$$\begin{aligned} \mathrm{FDI}_{ijt} = \; & \beta_0 + \beta_1 \mathrm{GDPsum}_{ijt} + \beta_2 (\mathrm{GDP}dif_{ijt})^2 + \beta_3 \mathrm{SKILL}dif_{ijt} \\ & + \beta_4 [\mathrm{GDP}dif_{ijt} \times \mathrm{SKILL}dif_{ijt}] + \mathbf{mc}_{ijt}\gamma + \lambda_j + u_{ijt} \end{aligned} \tag{20.16}$$

The dependent variable (FDI_{ijt}) is the real volume of production (sales) by manufacturing affiliates in each host country j that are majority owned by parents in domestic country i. The variable $\mathrm{GDP}sum_{ijt}$ is the bilateral sum of real GDP levels at home and abroad (the joint market size) whereas $(\mathrm{GDP}dif_{ijt})^2$ is the squared difference in real GDP between home and foreign countries. Then, $\mathrm{SKILL}dif_{ijt}$ stands for the difference in skilled-labor abundance in both countries. Note that the variable $\mathrm{GDP}dif_{ijt} \times \mathrm{SKILL}dif_{ijt}$ captures the fact that affiliates sales are higher when the home country has a small size and is skilled-labour-abundant. Finally, $\mathbf{mc}_{ijt}$ is a vector of multinationalization cost variables such as the perceived costs of investing in, and exporting to, the host country as well as the perceived trade costs in exporting to the parent country.

A fixed effect (λ_j) is introduced for each foreign country. The results with country-pair dummies are not reproduced in the paper. As expected, outward investment increases with the joint market size, the convergence in GDP between the parent country and any host country and the abundance in skilled workers of the parent nation. Moreover, when the country fixed effects are introduced, the difference in skill endowments has a smaller role but remains significant while the other variables keep the same impact.

These results seem to offer direct support for the KC model and to reject the horizontal model. Indeed, the last model predicts that absolute skill difference is negatively related to affiliate sales (Markusen and Venables (2000)). In contrast, in KC model, the production of foreign affiliates grows when the difference in skilled-labour abundance declines. However, Carr, Markusen and Maskus (2001) estimate a pooled coefficient on a difference term that takes both positive and negative values. This introduces a substractive linear constraint which can lead to a sign reversal in the pooled or restricted coefficient. Indeed, when the difference is negative (resp., positive), the rise in differences implies a convergence (resp., divergence) in skill-labour endowments. From the same database, Blonigen, Davies and Head (2003) exactly replicate the analysis of Carr et al. except that they consider the absolute values of skill difference. In this way, these variables are always decreasing in skill similarity. With this correct specification, Blonigen et al. obtain coefficient signs that support the horizontal model. This result suggests that the preponderance of multinational activity in developed countries is horizontal in nature.

20.4.1.3 Exports and FDI: Substitutes or Complements

Another way to determine the preponderance of horizontal FDI is to test whether FDI and exports are substitutes or complements. Our analysis in Sect. 20.2 suggests that substitution is the expected relationship under horizontal investments. This result arises from the fact that this model focuses on trade in final goods. When intermediate goods are introduced, foreign investment and export may simultaneously increase or decrease. Indeed, the rise in the production of affiliates induces an increase in imported inputs from the home country which corresponds to intra-firm trade when they come from parent companies (see for example Feinberg and Keane (2006) Hanson, Mataloni and Slaughter (2005)). Then, sales abroad of final goods and exports of intermediate goods can be complements. Several studies have examined the empirical relationship between production abroad and exports.

From a panel of Japanese firms over time, Head and Ries (2001) show that overseas investment and exports are complements. This result is obtained for the entire sample and by controlling for fixed firm effects. However, when the sample concerns the large assemblers that are not vertically integrated, the production of plants located abroad and the exports are substitutes. By using data published by the BEA which varies by country and over time (between 1977 and 1994), Clausing (2000) finds evidence that US multinationals activity and US exports are complements. Indeed, by using a gravity equation specification of trade, the author shows that a rise in affiliate local sales net of the value of imports from the US parent company increases the US exports. This result is robust when country-specific effects are controlled for.

It is clear that the relationship between foreign investments and exports depends on the level of aggregation of data. The studies using firm-level data underestimate the complementary effect since firms may purchase a number of inputs from independent suppliers that are set up in their domestic country. At the opposite, when

data are not disaggregated, the complementary effect is overestimated. Swenson (2004) examines how the change in US imports of product k from country i is related to changes in FDI stocks measured at three aggregation levels: product k (3-digit), 2-digit industry which produces k and overall manufacturing. By controlling for endogeneity, the empirical analysis reveals that US imports and foreign investment in US are substitutes at the product-level while, at the overall manufacturing level, they are complements. Note that any nation or industry fixed effects drop out from the estimating equation.

Finally, Egger (2001) proposes a dynamic treatment of the bilateral economic relationship, which would allow a useful distinction between short-run and long-run relationships. The analysis is based on a dynamic bivariate panel framework. The data cover the period 1986–1996 for bilateral relationships between the 15 EU members. The empirical model is given by:

$$\begin{aligned} dEX_{ijt} &= \alpha_0 + \alpha_1 dEX_{ijt-1} + \alpha_2 d\mathrm{FDI}_{ijt-1} + \alpha_3 Z_{ijt} + \rho_t + \varepsilon_{ijt} \\ d\mathrm{FDI}_{ijt} &= \beta_0 + \beta_1 dEX_{ijt-1} + \beta_2 d\mathrm{FDI}_{ijt-1} + \beta_3 Z_{ijt} + \gamma_t + u_{ijt} \end{aligned} \tag{20.17}$$

where dEX_{ijt} and $d\mathrm{FDI}_{ijt}$ are first differences of exports and stocks of outward FDI from country i to country j at period t, respectively. The use of first differences as well as the Hansen (1982) two-step generalized methods of moments controls for the correlation between lagged endogenous regressors and the error term. Note that exports is included in the FDI equation as a lagged variable. The explanation offered by the author is as follows: before setting up a plant in a country to serve this market, firms look at their export performance. Then, Z_{ijt} is a vector of variables similar to (20.15) and (20.16). Finally, ρ_t and γ_t are time-specific fixed effects. These effects take into account business cycles affecting Europe as a whole. The estimated coefficients of the lagged endogenous variables are significant suggesting that adjustment costs play is a major role in FDI and exports. The estimation results indicate also that outward FDI does not influence exports in the short-run, and vice-versa. Consequently, it is difficult to reach a clear conclusion on the complementary or substitutive nature of FDI and exports.

20.4.1.4 Exports and FDI: The Role of Distance

Among the key determinants of the decision to produce abroad, some variables such as distance and sunk costs do not vary over the time. However, using first differences or within transformation does not permit to measure the impact of all time invariant factors.[8] In addition, these explanatory variables are likely to be correlated with the time effect. As a result, the Hausman–Taylor model should allow for testing the role of the distance in FDI and using time effects. The difficulty arises from the choice of variables which are considered as doubly exogenous (not correlated with the unobserved effects) and as singly exogenous (correlated with

[8] Note that Carr, Markusen and Maskus (2001) do not take into account this problem when they estimate the model (20.16).

the unobserved effects). Although this econometric issue is very important, empirical papers presented in Sect. 20.4.1. do not take account of this bias. There are few papers using the Hausman–Taylor model to study the determinants of FDI. Egger and Pfaffermayr (2004a) is a notable exception.[9] Their data concern FDI from the US and Germany to other countries between 1989 and 1999. By controlling for (fixed) time effects and (random) industry-country pair effects, the authors find that distance has a significant and positive impact on outward FDI and that exports and outward FDI are complementary in the US and (weakly) substitutes in Germany. Moreover, an over-identification test suggests that distance and relative factor endowments are singly exogenous.

20.4.2 Testing the Role of Trade Policy in FDI

Almost all empirical contributions reviewed in the previous subsection consider that tariff-jumping is an important motive for FDI. The role of tariff barriers in the decision to produce abroad has also received specific attention from several empirical analysis. These studies are important since they test the ability of policy makers to influence international trade and FDI. The recent interest in the impact of trade policy on the decision to produce abroad arises also from important reductions in tariffs, quota and voluntary export restraints (VERs) and from an increasing number of countries with anti-dumping laws, because of numerous multilateral trade agreements (see Baccheta and Bora (2001) and Blonigen and Prusa (2003)). Consequently, it is not surprising that the tariff-jumping FDI analysis has concerned three main aspects: (i) anti-dumping (AD) policies; (ii) the threat of a protectionist policy (the so-called quid pro quo FDI hypothesis); and (iii) the transition periods of trade liberalization. Again, the use of panel data econometrics is crucial in these three domains. Indeed, testing the role of (i), (ii) and (iii) requires data that vary over time and information at product/firm level.

20.4.2.1 Effects of Anti-dumping Laws

As stated by Blonigen and Prusa (2003), *Since 1980, GATT/WTO members have filed more complaints under the AD statute than under all other trade laws combined.*... So, among other related questions, a growing number of empirical works using panel data study the effects of antidumping actions on FDI (these are mostly oriented toward Japanese firms).

By using a panel of 7 countries (6 EU members plus the US) over the period 1980–1991, Barell and Pain (1999) estimate a model which relates Japanese direct investment flows in country j at time t to a variable[10] denoting the "discounted stock" of anti-dumping cases (SAD_{jt}) in the EU or in the US, where

[9] More precisely, a seemingly unrelated regression Hausman–Taylor model is considered because they specify a system of two equations, exports and outward FDI, as in Egger (2001).

[10] In addition to several other regressors, such as market size and relative labour cost.

$$\mathrm{SAD}_{jt} \equiv \mathrm{AD}_{jt} + \sum_i (\mathrm{AD}_{jt-i})/i. \tag{20.18}$$

In this way, past anti-dumping actions may have a persistent, but progressively weaker, effect on Japanese FDI. The model is estimated by using the within estimator. The main result is that the level of AD has a positive effect on FDI. However, the authors use a very aggregated and quite short panel data ($N = 7$ and $T = 12$), even if they test carefully for the presence of heteroscedasticity and serial correlation of errors (see their Appendix A).

A more convincing analysis is provided by Blonigen (2002).[11] He first observes that: *In August 1993, Eastman Kodak Company filed a US antidumping petition against US imports of photographic paper originating from plants owned by Fuji Photo Film in Japan and the Netherlands.... While this led to an ensuing suspension agreement that led to substantially lower imports for a brief period, Fuji soon located a photographic paper manufacturing plant to the United States....* AD duties may result from a complex mechanism which requires the use of very disaggregated data. They are observed by all firms, they may change over time when the foreign firm modifies its dumping behavior (it may obtain refunds of AD duties in some cases) or if the US Department of Commerce changes the way it fixes AD duties. For this reason, Blonigen uses a panel data including firm and product combinations involved in US anti-dumping investigations from 1980 through 1990. By using a probit model, the author evaluates the probability for a Japanese firm, subject to anti-dumping duties, of locating its production for a given product in the US. From a technical point of view, the model used is a pooled probit which includes industry dummies in order to control for unobserved industry characteristics. The main result is that AD duties have a significant but small effect on FDI probability. Moreover, this effect is stronger when the firm has previous multinational production experience.

One interesting variant of Blonigen's approach may be found in Girma, Greenaway and Wakelin (2002). The authors introduce a time dimension in the panel and apply a different estimation strategy. Their basic model explains the presence (measured in terms of employment or fixed assets) of Japanese firms in the UK by a set of explanatory variables including the cumulated number of anti-dumping cases against Japanese firms measured as in (20.18). In this way, past anti-dumping actions may have a persistent, but progressively weaker, effect on Japanese FDI. The panel consists of 223 sectors observed over 1988–1996. The variables are constructed by aggregation of firms data. This permits a better evaluation of the tariffs and cumulative anti-dumping variables, as well as accounting for their time variabilities. However, for 146 industries, the dependent variable is equal to zero. So, the authors use Heckman's two-step estimation method. In the first stage, they estimate the probability of having Japanese FDI in the sector by using a probit model. In the second stage, they restrain their sample to the sectors with strictly

[11] Belderbos and Sleuwaegen (1998) follow broadly the same approach with data on Japanese FDI in the EU. They use a panel consisting of 131 firms and 345 (electronic) products, and so observed at a very disaggregated level. The authors confirm that VERs, antidumping actions and tariffs favor Japanese FDI and have a negative effect on firm-level exports to Europe.

positive FDI and explain the level of FDI in these sectors by unit labor costs and the cumulative number of anti-dumping cases. Additionally, in order to control for a selectivity bias, the inverse Mills ratios estimated at the first stage are introduced in the second stage. The model also includes time dummy variables which capture some UK business cycle effects. The main result is that Japanese FDI in the UK depends significantly on anti-dumping actions, and, to a more limited extent, on VERs and tariff barriers.

20.4.2.2 Effects of the Threat of Protectionism

In the literature on the quid pro quo FDI hypothesis (see Bhagwati, Dinopoulos and Wong (1992) and Grossman and Helpman (1996)) FDI may be caused by the threat of protectionism, and not only by actual protectionism as in the tariff-jumping analysis. Foreign investment may be used by international corporations as an instrument to defuse a possible protectionist action. In this case, when a firm establishes an overseas local production unit and creates jobs, the host country has less incentives to adopt protectionist measures.

Testing the quid pro quo hypothesis is difficult because the threat of protection is not observed and must be distinguished from actual protection.[12] Blonigen and Feenstra (1997) have proposed a solution using a less aggregated panel dataset of Japanese FDI in the US across 4-digit manufacturing industries from 1981 to 1988. First, they define the threat of protection in industry i and year $t-1$ as a latent variable (Z^*) defined by

$$Z^*_{it-1} = \mathbf{w}_{it-1}\gamma + \eta_{it-1} \tag{20.19}$$

i.e. it relies on a set of variables **w** (including real Japanese import growth, US real GNP growth). Now, consider that we observe at time $t-1$ if an US anti-dumping action is being engaged $(Z_{it-1}=1)$ or not $(Z_{it-1}=0)$ by the administration against Japanese firms in a given industry. Suppose also that,

$$\begin{cases} Z_{it-1} = 1 & \text{when } Z^*_{it} > 0 \\ Z_{it-1} = 0 & \text{when } Z^*_{it} < 0 \end{cases} \tag{20.20}$$

Hence, a US anti-dumping action (Z_{it-1}) at time $t-1$ is an indicator of the threat of protection (Z^*_{it}) at time t. In a first step, model (20.19) is estimated as a pooled probit model (a random effects model would be probably a better solution) which allows the authors to compute the predicted probability of protection $\widehat{Z}_{it-1}$. Finally, this variable is introduced in the Japanese FDI equation:

$$\text{FDI}_{it} = \mathbf{x}_{it}\beta + \widehat{Z}_{it-1}\delta + \varepsilon_{it} \tag{20.21}$$

[12] A previous attempt to evaluate quid pro quo FDI was done by Ray (1991). Unfortunatly, the analysis is conducted at industry level, probably inappropriate because trade protection is more often product-specific.

where $\mathbf{x}_{it}$ contains a variable measuring actual protection in addition to $\widehat{Z}$. Then, this specification evaluates separately actual protection and threat effects. Equation (20.21) can be estimated in a convergent way by OLS if η_{it-1} is independent of ε_{it}.[13] In fact, as FDI values are not systematically reported in their database (ITA), the authors choose to specify their dependent variable in the second equation as the discrete number of FDI occurrences in a 4-digit industry in year t. As a result, a random effects negative binomial specification[14] is adopted which is an extension of the Poisson model by introducing an individual unobserved effect in the conditional mean (each industry is assumed to be characterized by a specific propensity to do FDI). The main result is that Japanese FDI are highly sensitive not only to the actual anti-dumping measures but also to the threat of such measures.

Finally, notice that without the use of panel data (sectors/firms and time), it is probably impossible to split tariff-jumping and qui pro quo effects. Nevertheless, the random effects negative binomial specification requires that $\mathbf{x}_{it}$ to be strictly exogenous conditional on the unobserved effects. As R&D expenditures are included in the regressors, this may raise some problems. As suggested by Hausman, Hall and Griliches (1984), it would be useful to estimate a fixed effects negative binomial model which allows for dependence between $\mathbf{x}_i$ and the unobserved heterogeneity term.

20.4.2.3 Effects of Periods of Trade Liberalization

Another way to assess the impact of tariffs on FDI is to study how MNFs react during trade liberalization periods or when regional economic integration occurs. Over the past years, there has been an important increase in efforts among countries to achieve regional economic integration. Trade agreements largely differ on the degree of integration they imply: free-trade areas (NAFTA-1994, EFTA-1960); customs unions (Mercosur-1995), common markets (European Single Market-1992); or economic unions (Maastricht Treaty on the European Union-1998). Most studies on the relationship between regional integration and FDI have focused on the EU and NAFTA experiences.

Concerning NAFTA, an interesting analysis is due to Feinberg and Keane (2001). They analyze the effects of US and Canadian tariff reductions on the production location decisions of 701 majority-owned US-based MNF parents and their Canadian affiliates.[15] Their study has two main interesting features. First, data are observed at a firm level (and not as usually at a more aggregated one industry and/or country level) over a relatively large period (1983–1992) that includes both the Tokyo Round and the Canada-US Free Trade Agreement. Such panel data allow authors to examine the effects of tariff reductions on changes in MNF production-location and, at the same time, to control for time, firm and industry effects. Secondly, the

[13] See the discussion in Blonigen and Feenstra (1997), and especially footnote 10. See also Maddala (1983).

[14] See Hausman, Hall and Griliches (1984).

[15] They observe that US and Canadian tariffs dropped by approximately 62.5% from 1983 to 1992.

authors use a random coefficient approach.[16] The regression model is expressed as follows:

$$\begin{aligned} Y_{it} &= \beta_0 + (\beta_1 + \mu_{i1})CT_{it} + (\beta_2 + \mu_{i2})UT_{it} \\ &\quad + (\beta_3 + \tau_i)\text{TREND}_t + \beta_4 Z_{it} + \phi_i + \varepsilon_{it} \\ \text{with } \varepsilon_{it} &= \rho\varepsilon_{it-1} + \eta_{it} \end{aligned} \tag{20.22}$$

where $\mu_{i1} \rightsquigarrow N(0, \sigma^2_{\mu_1})$, $\mu_{i2} \rightsquigarrow N(0, \sigma^2_{\mu_2})$ and $\tau_i \rightsquigarrow N(0, \sigma^2_{\tau})$. The variable Y_{it} is defined in different ways, for instance exports from Canadian affiliate i to its US parent or exports from US parent i to its Canadian affiliate. CT_{it} and UT_{it} are respectively Canadian and US tariffs in the industry to which firm i belongs at time t, and Z_{it} includes others exogenous variables like transport costs, relative factor costs, GDP for each country and manufacturing wages. Such a specification has several advantages. First, the μ_i's may capture across-firm heterogeneity in tariff responses whereas τ_i and ϕ_i control for heterogeneity in the time trend (business cycle) for the former and for unobserved time-invariant firm specific characteristics for the latter. Note also that this specification is quite parsimonious if we compare it to a fixed effects approach. Second, once the population mean for each β and the variance of the β_is are estimated, the authors construct estimates (a posteriori) of the individual firm β_i. Then they compute the mean of each β_i within several industries defined at the disaggregated 3-digit level. Lastly, they decompose the total variance of the firm-specific β_i between across- and within-industries. The main results are twofold. First, the effect of Canadian tariff reductions on US parent exports to Canadian affiliates is very low (a 1% reduction in the Canadian tariffs increases US parent sales to Canadian affiliates by 1.6% on average, and moreover, the coefficient is significant only at the 20% level). At the same time, reductions in the US tariffs imply a greater Canadian affiliate production for sales into the US. Hence, *trade liberalization appears to have been trade-creating...* and does not induce a "hollowing out" of Canadian manufacturing. Second, within-industry (firm) effects explain more than 75% of the variance in the random tariff coefficient. So, firms' response to a change in the tariff depends heavily on unobserved firm characteristics (technology and organization). Industry characteristics (scale economies and product differentiation) are not a major determinant of the pattern of adjustment even if the industry is narrowly defined. Nevertheless, as noted by the authors themselves, the random coefficients specification may not be adequate if adjustment costs in production are high and if a negative cross-sectional correlation exists, at the preliberalization period, between tariffs and trade flows.[17] In this case, a fixed-effects Tobit model may be a solution, but its estimation is computationally more difficult.

[16] In fact, the model is a random effects tobit model, estimated by ML, because some Y_{it} are equal to zero, when, for instance, affiliates produce and sell all their production in Canada.

[17] See footnote 16, p. 127 in Feinberg and Keane (2001) for more details.

Concerning the EU, there exist many studies on the impact of the various stages of the integration process on FDI.[18] Typically, regional trading groups, currency unions... are captured by dummy variables. More recent works on this topic rely on two further considerations. First, an integration process takes time to be implemented and absorbed by the economies. So, integration effects must be modeled in a dynamic way, in order to distinguish between short-run and long-run effects and between announcement and post integration effects. Second, as it is necessary to control for many unobserved factors (host and home countries, time, integration phases effects), many dummies have to be introduced in the model, which can lead to a serious loss of degrees of freedom and/or multicollinearity problems. Several recent papers deal with these issues.[19]

Egger and Pfaffermayr (2004-b) try to isolate the impact on FDI of three EU integration phases: the Single Market Program, the 1995 enlargement of the EU and the Agreements between the EU and the Eastern European countries. They use a FDI gravity model with bilateral and time effects in which they add 20 bilateral integration group effects (e.g. EU 12, EFTA, rest of the World, CEEC) interacted with all the three phase dummies (1986–1992, 1993–1994 and 1995–1998) that gives 60 integration dummies. The model, estimated by the within estimator[20], may be defined as:

$$\text{FDI}_{ijt} = \mathbf{x}_{ijt}\beta + \alpha + \lambda_t + \delta_{ij} + \pi_{kp} + u_{ijt} \tag{20.23}$$

where $p\ (= 1,2,3)$ represents the integration phase and $k\ (= 1,\ldots,20)$ the country group. The estimation period is 1986–1998 and the unbalanced panel contains 3642 observations (with 13 home and 55 host countries). The main conclusion is that the integration effects on FDI are substantial and positive, but largely anticipated by the countries. Once the integration process is officially completed, regional integration has no more effects on FDI. However, the difference-in-differences estimator does not eliminate factors evolving differently over time between countries. So, if unobserved heterogeneity remains in the data, omitted variable bias may be a real problem. Moreover, as the number of countries is not "large" in this work, correct inference may be complicated (see Wooldridge (2003)).

In a very detailed work, Nicoletti, Golub, Hajkova, Mirza and Yoo (2003) use new structural policy indicators constructed by the OECD to estimate the impact of various trade policies on trade and FDI. Among many factors (FDI restrictions, bilateral tariffs and non-tariff protection), they study the role of belonging to a free trade area on FDI. They estimate two bilateral equations of FDI (one for outward stocks, one for outflows). The general model is

[18] One of the first attempts is given by Brenton, Di Mauro and Lucke (1999) who unfortunatly do not use the panel dimension of their data.

[19] There is a growing empirical literature on this subject, e.g. Girma (2002), Mold (2003), Altomonte and Guagliano (2003), Bevan and Estrin (2004), Carstensen and Toubal (2004) and Yeyati, Stein and Daude (2003) among others. More recent papers take into account the endogeneity of free trade agreements (Baier and Bergstrand (2007)) by of estimating a model on panel data with IV and control-function techniques.

[20] Also called, in this context, the difference-in-differences estimator.

$$\begin{aligned} \mathrm{FDI}_{ijt} = &\sum\nolimits_x \beta_x \mathbf{X}_{ijt} + \sum\nolimits_c \beta_c \mathbf{C}_{it} + \sum\nolimits_p \beta_p \mathbf{P}_{jt} \\ &+ \alpha_i + \alpha_j + \alpha_{it} + \alpha_{ij} + \alpha_{jt} + u_{ijt} \end{aligned} \tag{20.24}$$

where i (resp., j) represents the home (resp., host) country, FDI_{ijt} is the log of bilateral FDI outward stocks or flows at time t, X_{ijt} are country-partner pair specific variables, C_{it} are country specific variables, and P_{jt} are partner pair specific variables. As in Egger and Pfaffermayr (2004-b), the model contains many dummies in order to control for observed and unobserved factors relative to time and (host and home) countries. Nevertheless, Nicoletti et al. adopt a different estimation strategy. Host-specific and home-specific effects are eliminated by using "transformed least squares", i.e., by expressing the data as deviations from the average home country or the average host country. In this way, all home and host specific dummies are removed from the model. They use OECD data described in the previous section, so potentially $28 \times 27 \times 21 = 15876$ observations are available, but, due to numerous missing values, only about 4500 are used in the estimations. The main conclusion is that participation in free-trade agreements has had significant quantitative effects on FDI, particularly within the EU. For instance, they estimate the increase in FDI stocks to be up to 100% for Czech Republic, Hungary and Poland between 1990 and 2004.

As shown by the two previous works, dynamic aspects (e.g. anticipations) seem to play a major role when one tries to assert effects of regional integration on FDI.

20.4.3 Testing the Relationship Between FDI and Exchange Rate

In the second half of the 1980s the value of the yen increased while the dollar experienced a sharp depreciation. This phenomenon could explain why Japanese FDI increased rapidly in the US during this period. However, the relationship between the exchange rate and FDI is not evident. Under the assumption of a perfect international credit market, firms have the same advantage/disadvantage to purchase any particular asset abroad or at home. In other words, entrepreneurs are able to borrow at the same opportunity cost whatever their location and their nationality. Consequently, the variations of exchange rates do not affect the structure of the private capital account of countries' balance of payments between portfolio investment and FDI. Since the beginning of the 1990s, imperfections in the capital market have become the main argument used in the literature to justify why it is necessary to revisit the relationship between exchange rates and FDI. This question is discussed in the next subsection. We will report empirical studies that highlight the role of imperfections on the product market, on the one hand, and, the volatility of the exchange rate, on the other hand, in the relationship between exchange rate and foreign investments.

20.4.3.1 Role of Imperfections on Capital Markets

Two types of capital market imperfections play a key role in the relationship between the decision to produce abroad and the exchange rate: the existence of asymmetric information and the capacity of banks to grant loans.

Froot and Stein (1991) propose an adverse selection model where there exists asymmetric information between lenders and borrowers/firms about the future profit from an investment project. Moreover, the creditors incur a monitoring cost if they want to observe the profit realized by the borrowers. This monitoring cost is what causes external resources to be more expensive than internal resources and explains why firms do not finance the whole of their investment by loans. The investment project concerns the purchase of a domestic firm either by another domestic company or by a foreign multinational firm through a bidding process. As the domestic currency experiences a real depreciation, the self-financing capacity of the MNF grows relative to that of the other domestic bidder, so that (ceteris paribus) the MNF increases its probability of winning the auction. From this analysis, the link between real exchange rate and FDI is obvious.[21]

To verify the validity of these different theoretical arguments, Froot and Stein (1991) use annual panel data coming from the International Trade Administration (ITA) of the US Trade Department and for the period 1977–1987. From this database, the authors examine whether the wealth effect may be differentiated across industries or across different types of FDI. Indeed, the US FDI inflows are disaggregated by source country, and by industry as well as by type of purchases/transactions (plant acquisition or expansion, merger and acquisition, joint-ventures). Results suggest that the real exchange rate has not the same effect on the different parts of the total foreign capital inflows into the US. The dollar variations only have a significant effect on inward direct investments, as expected. The estimates of the real exchange rate effects seem more convincing at the level of the different types of FDI transactions. The exchange rate has a statistically significant impact with the right sign on FDI associated with mergers and acquisition operations.[22]

Nevertheless, as the model is estimated by pooled OLS, it does not take into account individual or time effects. This limit is important since no other variables are introduced in the model which control for the alternative explanations of FDI (such as distance, trade costs).

This criticism has been removed by Klein and Rosengren (1994). They adopt the approach retained in Froot and Stein (1991) but consider a fixed-effects specification to take into account the heterogeneity between the source countries. Moreover, considering that the ITA data used by Froot and Stein (1991) are not necessarily comprehensive, Klein and Rosengren prefer to complete their empirical analysis

[21] This analysis does not hold for other types of inward investment such as foreign investment in Treasury securities or in corporate stocks and bonds. For these portfolio investments, the monitoring costs are expected to be small and, thus, uncorrelated with the real exchange rate.

[22] The exchange rate has also a significant impact on joint-ventures and new plant FDI. This last result is problematic since the bidding approach developed by authors does not really concern these types of transaction.

by using the BEA measure of FDI although this includes foreign acquisitions of existing American-target firms and the establishment of new plants by MNF.[23] The available sample for the BEA series is over the 1979–1991 period while the ITA annual data concern 1977–1987. As in Froot and Stein, the real exchange rate has always a statistically significant impact with the right sign not only on FDI, as a whole, but also on foreign mergers and acquisitions operations. Note that in accordance with the theoretical conclusions of Froot and Stein, the effect is lower on FDI than on mergers and acquisitions. On the other hand, as a log–log specification is only used by Klein and Rosengren, it is difficult to know whether the estimation of the fixed-effects model really modifies the values of the parameters compared to the pooled OLS estimation. In order to control for alternative explanations for FDI, Klein and Rosengren introduce as a regressor in their model the relative-labor-cost between the US and the source countries. They find that the wealth effect is always at work for FDI through mergers and acquisitions in the US while the relative-labor-cost has no impact. Then, these different results suggest the empirical validity of the conclusions drawn by Froot and Stein. However, the potential correlation between the disturbances of the models relative to the different type of FDI is not taken into account through, for example, SUR estimation.

However, during the mid-1990s, Japanese FDI fell whereas the yen appreciated significantly. To explain this feature, Klein, Peek and Rosengren (2002) focus on the role played by the financial intermediation. In a country where the relationships between firms and banks are very close, the financial intermediation is dominant. In this context, firms' ability to engage in FDI is influenced by the capacity of banks to grant loans. It is the relative access to credit (RAC) hypothesis. In the 1990s, the Japanese bank sector experienced a collapse causing Japanese firms to be constrained in the financing of their investment projects. Thus, the value of Japanese FDI as a share of total inward US FDI reached a peak of 30% in 1990 and then declined during the following years by only 1% of total inward US FDI by 1998. It is the validity of the RAC hypothesis that is tested by Klein, Peek and Rosengren (2002).

A database is constructed by Klein et al. from firm-level FDI ITA over the period 1987–1994. They use the number of FDI projects since the amount of FDI is not systematically available. Moreover, Japanese firm characteristics (size, profitability, market value and industry) come from the Pacific-Basin Capital Markets Databases. From the Japan Company Handbook, are identified the 11 primary (first referenced) banks of the Japanese firms included in the sample. During the sample period, few Japanese firms change their primary bank. In order to obtain an independent/objective evaluation of the banks' financial health and their evolution over the sample period, Klein, Peek and Rosengren (2002) use the time series of Moody's long-term deposit ratings. Thus, the authors exploit the time-heterogeneity between banks although all of them experienced a downgrade in their Moody's ratings during the last years of the period. The empirical model is represented as follows:

[23] Despite the difference in definition of FDI, the correlation between the BEA measure and the ITA measure is quite high (0.86). This result confirms the preponderance of mergers and acquisitions in the US inward FDI.

$$\text{RATEFDI}_{it} = \alpha_0 + \alpha_1 \text{DPROFIT}_{it-1} + \mathbf{drating}_{it-1}\beta + \mathbf{dmacro}_{it-1}\gamma + \varepsilon_{it} \tag{20.25}$$

where $\varepsilon_{it} \rightsquigarrow N(0, \sigma^2)$. In this specification, the dependant variable, RATEFDI_{it}, is the variation rate in the number of FDI projects toward the US, financed by Japanese primary bank i during year t. $\mathbf{drating}_{it-1}$ contains two measures of changes of Moody's long-term deposit ratings for the Japanese main banks. A first dummy variable takes the value 1 if the bank i has a change in its rating, during year $t-1$, and 0 otherwise. Its effect on FDI is a priori negative. A second dummy variable is introduced. Its coefficient is also expected to be negative because this variable takes the value 1 when there are two or more downgrades and 0 otherwise. DPROFIT_{it-1} corresponds to the variation of the profit sum of the firms associated with bank i. This variable measuring the change in the health of firms is assumed to favor FDI. $\mathbf{dmacro}_{it-1}$ contains a set of three macroeconomic variables intended to control for differences in wealth and economic activity between Japan and the US. The first variable is introduced to control the variation of wealth between both countries in the spirit of Froot and Stein (1991). The impact of this variable on FDI should be positive. The change in the US unemployment rate and the change in the Japanese job-offers-to-applicants ratio are used to control for the macroeconomic business cycle in both countries. The effect of these two variables on Japanese FDI to United States is assumed negative.

The coefficients estimated are in accordance with the expectations. Thus, the multiple-level (single-level) downgrade of a bank during the period causes a 70% (30%) reduction in the number of Japanese FDI projects that use this bank as their main lender. On the other hand, the wealth effect is not statistically significant, weakening the argument developed by Froot and Stein (1991). Then, the decreasing number of the Japanese FDI projects in the US over the 1990s seems to be explained by the collapse of the Japanese banking sector rather than by the loss of competitiveness of the Japanese Firms. Moreover, to show the robustness of their estimates, Klein, Peek and Rosengren (2002) provide estimates from two restricted samples including only multiple-year FDI firms or banks financing the most FDI projects. In fact, the individual effects, which are likely to be correlated with some regressors, have not been removed by the variable transformation used. Indeed, the first difference transformation is not applied to the explained variable since DFDI_{it} is a rate. Therefore, the estimation method is pooled OLS which is biased and inconsistent.

20.4.3.2 Role of Imperfection on Product Markets

The relationship between FDI and the exchange rate can also be explained by imperfections in the product market. Blonigen (1997) establishes three conditions for the existence of a specific relationship between the (real) exchange rate and FDI. First, the opportunity to purchase a target firm which owns a specific asset. The transfer of this specific asset is realized at a low cost between different facilities whatever their nationality. The target firm may be bought either by a (US) domestic firm or by

a (Japanese) foreign one. Second the domestic and foreign markets are segmented. This market imperfection challenges the law of one price and price adjustments that could compensate for a change in the nominal exchange rate. Third, the access to the foreign market must be limited to the domestic firm. Otherwise, both acquiring rivals would have the same return on the specific asset abroad. Then, the domestic firm knows entry barriers on the foreign market. These three conditions being verified, a real depreciation of the domestic currency (the US dollar) relative to the foreign one (the yen) leads to an increase in the surplus of the foreign firm. Consequently, the foreign firm has an incentive to make a higher bid than its rival for buying the target firm. The direct consequence is that a greater foreign acquisition of the US assets must be expected during a period of real dollar depreciation, other things being equal. This analysis exclusively concerns the inward FDI associated with mergers and acquisitions operations.

From a balanced panel of 361 industries both manufacturing and nonmanufacturing over the period 1975–1992, Blonigen (1997) analyzes the positive relationship between the number of Japanese acquisitions by industry and by year (NFA_{it}) into the US and the real exchange rate (RER_{it}) at industry level.[24] The specification has the following form:

$$\Pr(\text{NFA}_{it}) = f(\text{RER}_{it}, \Omega_{it}, \Psi_{it}) \tag{20.26}$$

where Ω_{it} includes variables having an important role in this approach: (i) the number of acquisitions of US target firms by other US firms (proxying the supply of specific assets on the US market); (ii) the share of Japanese value added in each industry (for the US market penetration of Japanese firm); (iii) the annual real growth of Japanese GDP (a proxy for Japanese demand for specific assets); (iv) the annual growth in the Tokyo Stock Price index (its effect is assumed positive). This variable is used as a proxy for the outgrowth of the speculative "bubble" economy of Japan in the late 1980s and early 1990s. The variables included in Ψ_{it} must control for other explanations found in the traditional literature on FDI, analyzed previously.

The data on the number of foreign acquisitions are typical of count data which can vary from zero to several or even many, for some industries. The negative binomial model (Hausman, Hall and Griliches (1984)) is used for estimation and represents a generalization of the Poisson distribution with an additional parameter allowing the variance to exceed the mean. Indeed, for the manufacturing and nonmanufacturing sectors over the 1975–1992 period, the number of Japanese acquisitions (NFA) is ranged from 0 to 89, with a mean of 16 and a standard deviation of 20. Beyond this problem of overdispersion (or underdispersion), the model includes individual fixed or random effects to take into account the cross-sectional heterogeneity. From this model and under the assumption of fixed effects, $\Pr(\text{NFA}_{it})$ can be written as,

$$\Pr(\text{NFA}_{it}) = \frac{\Gamma(\lambda_{it} + \text{NFA}_{it})}{\Gamma(\lambda_{it}\Gamma(\text{NFA}_{it} + 1)} \left(\frac{\theta_i}{1+\theta_i}\right)^{\lambda_{it}} \left(\frac{1}{1+\theta_i}\right)^{\text{NFA}_{it}} \tag{20.27}$$

[24] The dollar value for acquisitions is not retained as a dependent variable since it is missing for over one-third of the observations.

where Γ is the gamma function. The parameter θ_i is the individual effect while λ_{it} depends on the covariates by the following function:

$$\ln \lambda_{it} = \alpha \mathrm{RER}_{it} + \Omega_{it}\beta + \Psi_{it}\gamma \tag{20.28}$$

Under the assumption that NFA_{it} are independent over time, $\sum_t \mathrm{NFA}_{it}$ also has a negative binomial distribution with parameters θ_i and $\sum_t \lambda_{it}$. In a context of random effects now, to permit a tractable negative binomial mass function, $\theta_i/(1+\theta_i)$ is assumed to be distributed as a beta random variable with shape parameters (a, b). The model with fixed or random effects can be estimated via the maximum-likelihood method Greene (2004).

The estimates of both types of models support the main hypotheses formulated by Blonigen (1997). In other words, a real appreciation of the yen relative to the dollar leads to an increase in the number of Japanese acquisitions in the US. This result holds for manufacturing industries rather than for nonmanufacturing and for the industries with high levels of R&D. Moreover, the parameters associated with the US supply and Japanese demand of specific assets have the positive expected sign. On the other hand, the alternative explanations of Japanese FDI based on the US tariff-jumping and the Japanese speculative bubble are unsatisfactory. Therefore, these different results exhibit in accordance with Froot and Stein (1991) a wealth effect. However, this effect is mainly present in industries where specific assets like innovation are present and it concerns foreign mergers and acquisitions, the most important component of FDI. Then, this analysis goes further into the specific relationship between FDI and exchange rate.[25]

However, no proxies are introduced to estimate the level of entry barriers in the Japanese markets for the US firms while this is one hypothesis among the most important of this original approach. Moreover, the estimates from random and fixed effects models are both reported but they are not compared using a Hausman test to indicate whether the industry-specific effects are correlated with the regressors. Without results of this specification test, it is difficult to deduce the appropriateness of both models. Another important issue is how to introduce the fixed effects in the negative binomial model. In this case, the fixed effects are conditioned out of the likelihood function (see Allison and Waterman (2002)).

20.4.3.3 Role of the Exchange Rate Volatility

Since the 1970s and the end of generalized system of fixed exchange rates, economists have also devoted much attention to the effects of exchange rate volatility on FDI. The effect of volatility of the real exchange rate on FDI depends on whether firms may choose to export or to invest abroad, on their behavior towards risk and on their expectations about the future profits from FDI.

[25] These results are broadly supported by the empirical analysis of Guo and Trivedi (2002) that assigns the industries to high- and low-sensitivity FDI categories relative to the exchange rate movements.

Following Dixit (1989), Campa (1993) develops an option model to explain why the MNF make FDI or not (see also Altomonte and Pennings, 2003). In this framework, as opposed to the traditional theory of investments under uncertainty, the expected future profits of a MNF, assumed risk-neutral, take into account the exchange risk of entering the foreign market.[26] Assuming the future values of the nominal exchange rate are lognormally distributed with a variance σ that grows linearly with the time horizon. Then, even if a positive drift may lead to an appreciation of the foreign profits in domestic currency, at the same time the volatility in the future of the nominal exchange rate may be too important to discourage FDI. In fact, the MNF retains an option to enter the foreign market at any moment in time. This option has a price which is the sunk cost of entering the market through FDI. This sunk cost makes an irrevocable commitment of the MNF when it exercises the option. Moreover, the value of the option– equivalent to the value of the investment opportunity- is the expected present discounted value of future profits from serving the foreign market minus the amount of FDI realized at time t. Here, the decision to make FDI is equivalent to deciding at which time to exercise such an option. The MNF holds back for an extra period (e.g. stays out of the foreign market one more time) as long as the expected change on the option's value is higher than the expected return of the present time. From his theoretical model, Campa (1993) deduces some predictions about the effects of exchange rate volatility on FDI. The higher the exchange rate R, the rate of change μ and the uncertainty σ, the more valuable the option to enter is and the fewer events of entry observed. Furthermore, the lower the marginal cost MC and the cost to entry in the foreign market k, the higher the expectation of future profits from the activity abroad. Note that the marginal cost is not expressed in foreign currency since Campa (1993) limits his empirical analysis to FDI related to wholesale activities. Thus, the model concerns a MNF producing a good in the home country and selling it in a foreign market via a sales subsidiary. The verification of these different predictions allows the author to construct the following reduced form:

$$n^*_{it} = f(\mu, \sigma, e_{f/\$}, k, \mathrm{MC}) \tag{20.29}$$

where the explained variable, n^*_{it}, is the number of MNF that enter the US wholesale trade industry i in a given year t. The construction of regressors μ and σ depends on the MNF's expectations about the evolution of these two variables in the future. These two variables correspond to the average and the standard deviation, respectively, of the monthly change in the logarithm of the anticipated exchange rate. Two types of anticipation are considered: perfect and static expectations. The exchange rate level $e_{f/\$}$ is defined as the annual average of the exchange rate in units of domestic currency of the MNF per one unit US dollar, in the year of entry. The level of the sunk costs k is proxied by two variables. The first regressor is the ratio of fixed assets to net wealth of all US firms in an industry and the second is the ratio of media expenditures to turnover in each US industry. The marginal cost MC is proxied

[26] Note that the uncertainty can also concern the foreign production costs (see Brandao De Brito and De Mello Sampayo, 2002)

by the unit labor cost w since capital is assumed fixed, its cost being included in the entry cost k.

The database is a panel of 61 US wholesale trade industries, defined at the four-digit level, for the period 1981–1987. This sample contains a total of 768 entries of MNF in the US. Even though the MNF come from 35 different countries, the geographical concentration of origin countries is high. Thus, Japan, the UK, Germany, France and Canada account for almost 80% of the 768 foreign entries on the US markets. The dependant variable n_{it}^{*} is censored since it takes values from 0 to 40 with a large fraction of zeros. To take account of the truncated distribution of the dependant variable, Campa (1993) uses a Tobit estimation.

The estimates are in accordance with the predictions of the theoretical analysis. Thus, the estimated parameters have the right sign, except for μ. The uncertainty of the exchange rate has a significant negative effect on the number of MNF entering the US markets. The labor cost w is not significant in the estimates weakening the range of the theoretical model. Nevertheless, both proxies of the entry costs *Sunk* and *Adv* have significant negative coefficients while the level of the exchange rate R has a positive effect, as expected. However, this last result is the opposite of the conclusions of previous papers where the specific relationship between FDI and the level of exchange rate is explained through market imperfections. In fact, the wealth-effect argument developed by Froot and Stein (1991) is not valid in the present analysis. Indeed, this argument concerns only the FDI realized in manufacturing industries and it cannot be used for FDI in wholesale trade industries.

20.5 Some Recent Econometric Issues

The use of panel data presents specific features which introduce econometric complications. In this last section, we discuss two main problems that have been recently treated in the estimation of FDI models.

20.5.1 FDI, Panel Data and Spatial Econometrics

The first problem concerns the specification of the empirical model. Recent theoretical developments have stressed that a MNF may engage in FDI activities depending not only on home and host characteristics, but also depending upon the neighboring host's specificities. First, a multinational firm may use a host country as an export platform to other near markets for minimizing trade costs. Second, it may also split its production in several vertical units with respect to the relative factor costs between countries. By definition, an econometric bilateral model of FDI does not take into account the specificities of the neighboring host country. In order to control for the correlation between inward FDI of one country and FDI of its neighbors, we may use spatial panel data model estimation methods. Baltagi, Egger

and Pfaffermayr (2007) and Blonigen, Davies, Waddell and Naughton (2007) are two interesting studies on this topic. The first one analyzes US outward FDI stock in country-industry pairs (in 1989–1999) whereas the second one focuses on FDI from the US to 20 OECD countries (between 1980–2000). We can summarize their approach as follows (for simplicity, we present mainly the Blonigen et al. methodology, which is simpler – but less general – than Baltagi et al.). Consider that the data are sorted by time t (first sort key) and by host country j (second sort key), so we can omit the t index. For simplicity, we present the specification assuming that the panel is a balanced one ($N_t = N$),

$$\mathbf{fdi} = \mathbf{X} \times \beta + \rho \times \mathbf{W} \times \mathbf{fdi} + \mathbf{u} \tag{20.30}$$

where **fdi** and **u** are $N \times 1$ vectors (row j refers to the host country j) and **X** is a $N \times k$ matrix of regressors. $\rho \times \mathbf{W} \times \mathbf{fdi}$ reflects the spatial autoregression term, where $\mathbf{W}(N \times N)$ is a (row normalized) spatial lag weighting matrix (for each year t) (see Chap. 19 in this volume for the expression of **W**) where the components $(w(d_{ij}))$ are a weighted function depending upon the distance between country i and country j. According to Blonigen et al., a positive ρ means an agglomeration effect or vertical production organization while a negative ρ suggests FDI are used for export-platform reasons. Further, the spatial error term is defined as

$$\mathbf{u} = \rho' \mathbf{W}\mathbf{u} + \varepsilon$$

with $|\rho'| < 1$. Notice also that the Baltagi et al. specification is more general since it includes country-industry-pairs effects and also spatially weighted average of regressors. Errors are spatially correlated when $\rho' \neq 0$. However, recall that OLS estimators are still consistent but are inefficient. Finally, one interesting aspect of this model is that $\mathbf{W} \times \mathbf{fdi}$ is endogenous and correlated with **u**. To estimate this model, Blonigen et al. apply a maximum likelihood method while Baltagi et al. use the fixed and the random effects 2SLS estimator (using the second and third order spatial lags of the exogenous regressors as instruments).

In Blonigen et al. and Baltagi et al., estimations exhibit a significant spatial dependence, which is negative in the former article and positive in the latter article. In addition, spatial correlation of errors are only detected in Baltagi et al. Even though results are different in some respects, the estimation of a spatial panel data model of FDI is required in order to control for the correlation between the inward FDI of different neighboring countries.

20.5.2 Exchange Rate, Unit Roots and Cointegration

As many empirical analysis on FDI very often use non-stationary variables and models in levels, it is necessary to test for unit roots and cointegration in order to avoid spurious regressions. This is a particular problem in FDI-exchange-rate models while it is not treated in the empirical literature (see for example Froot and

Stein (1991) and Klein, Peek and Rosengren (2002)). Up to now, there have been few studies analyzing FDI determinants on panel data with this methodology which, in the context of countries panel data estimation, presents several specific features.

On the one hand, it is frequently advanced that panel-based unit root tests have higher power than unit root tests based on time series. Indeed**,** as countries' data are used for FDI analysis, panel data sets are sometimes characterized by large N and large T dimensions. An illustration of such tests is given by Hsiao and Hsiao (2004) who use an (unbalanced) panel data with only 5 countries observed between 1987 and 2002. They apply several tests (for example IPS and ADF-Fisher tests[27]) and conclude that FDI is stationary, the exchange rate, GNP and wage differential variables are not. Moreover, they show that the first differences of the non stationary series are I(0), so these variables are integrated of order one. These results seem to be frequently encountered even if the opposite conclusion is often obtained in the literature about the stationarity of FDI.[28]

On the other hand, given this result, the estimation of FDI determinants, using cross-countries data over time, raises several other issues. First, if the micro relationships are made of I(1) variables where each country has its own specific cointegrating relation, it is probably better to estimate the model on each country separately. Nevertheless, with too few annual periods, making inference in such a context may be difficult. Second, a solution to deal with spurious regressions is to take first differences of the I(1) variables and to apply usual panel data estimators in a framework of pooling with I(0) variables. Nonetheless, in this case, variables that are constant in the time dimension are removed from the model and a part of the long-run information is removed. For these reasons, De Santis, Anderton and Hijzen (2004) and Hsiao and Hsiao (2004) suggest proceeding in two steps. First, panel cointegration tests are applied in order to guard against the spurious regression problem. Once again, using panel data may improve the small sample properties of such tests, even if there is not general agreement on this point. From several tests (e.g. multivariate augmented Dickey–Fuller, Im–Pesaran and Shin tests), they reject the null hypothesis that the residuals of the panel regressions are I(1), i.e. they reject the null hypothesis of no cointegration in their panel data. Second, they must decide how to do estimation and inference in panel data cointegration models.[29] Among many possibilities (e.g. OLS, Mean group, FMOLS, Within estimators), the two studies choose to use the within estimator. Indeed, as the residuals of the within estimator are stationary, within estimates are probably not spurious.[30] Moreover, when T is large and N is moderately large, Phillips and Moon (1999) shows that the within estimator consistently estimates in many cases the long-run effects and has a limiting normal distribution.

[27] See Chap. 9 in this volume.

[28] See Brandão de Brito and Mello Sampayo (2004).

[29] Brandão de Brito and Mello Sampayo (2004) estimate the cointegration relationships and the error-correction mechanism equations for each country separately. As $T = 7$, it is not sure that such an approach is correct.

[30] As the De Santis, Anderton and Hijzen (2004) model is a dynamic one, they applied also the Arellano-Bond estimator on first differences and find similar results.

Hsiao and Hsiao (2004) study the determinants of FDI flows in China from Hong Kong, Japan, Taiwan, Korea and the US between 1987 and 2002 ($N = 5$ & $T = 14, 15$ or 16) whereas De Santis, Anderton and Hijzen (2004) analyze the factors influencing stocks of FDI in the US from eight Euro area countries (1980–2001), so that $N = 8$ and $T = 22$. Even though the data sets used are different, the specifications are somewhat closed and include some common regressors (lag FDI, exchange rate, GDP and wage differential). Limiting the presentation to the exchange rate, in both studies, a negative and significant relationship is found between the exchange rate and FDI. Nevertheless, as lag FDI is included in the regressors list, it is questionable that the Phillips and Moon (1999) results apply.

Acknowledgments We would like to thank Joe Byrne for helpful comments on an earlier draft.

References

Allison, P. D. and Waterman, R. (2002). Fixed-Effects Negative Binomial Regression Models. In Stolzenberg R. M. (ed.), *Sociological Methodology*. Oxford: Basic Blackwell, pp 247–265.

Altomonte, C. and Guagliano, C. (2003). Comparative Study of FDI in Central and Eastern Europe and the Mediterranean. *Economic Systems*, 27:223–246.

Altomonte, C. and Pennings, E. (2006). The Hazard Rate of Foreign Direct Investment: A Structural Estimation of a Real Option Model. *Oxford Bulletin of Economics and Statistic*s, 68(5):569–593.

Baccheta, M. and Bora, B. (2001). Post-Uruguay Round Market Access Barriers for Industrial Products. *Policy Issues in International Trade and Commodities*, Study Series No. 12, UNCTAD. New York and Geneva: United Nations.

Baier, S. C. and Bergstrand, H. (2007). Do Free Trade Agreements Actually Increase Members' International Trade? *Journal of International Economics*, 71(1):72–95.

Balestra, P. and Negassi, S. (1992). A Random Coefficient Simultaneous Equation System with an Application to Foreign Investment by French Firms. *Empirical Economics*, 17(1):202–220.

Baltagi, B., Egger, P. and Pfaffermayr, M. (2003). A Generalised Design for Bilateral Trade Flow Models. *Economics Letters*, 80(3):391–397.

Baltagi, B., Egger, P. and Pfaffermayr, M. (2007). Estimating Models of Complex FDI: Are There Third-Country Effects? *Journal of Econometrics*, 140(1):260–281.

Barell, R. and Pain, N. (1999). Trade Restraints an Japanese Direct Investment Flows. *European Economic Review*, 43(1):29–45.

Belderbos, R. and Sleuwaegen, L. (1998). Tariff Jumping DFI and Export Subsitution: Japanese Electronics Firms in Europe. *International Journal of Industrial Organization*, 16(4):601–638.

Bevan, A. and Estrin, S. (2004). The Determinants of Foreign Investment into European Transition Economies. *Journal of Comparative Economics,* 32:775–787.

Bhagwati, J. N., Dinopoulos, E. and Wong, K.-Y. (1992). Quid Pro Quo Foreign Investment. *American Economic Review*, 82(2):186–190.

Blonigen, B. A. (1997). Firm-Specific Assets and the Link Between Exchange Rates and Foreign Direct Investment. *American Economic Review*, 87(3):447–465.

Blonigen, B. A. (2002). Tariff-Jumping Antidumping Duties. *Journal of International Economics*, 57(1):31–49.

Blonigen, B. A., Davies, R. B. and Head, K. (2003). Estimating the Knowledge-Capital Model of the Multinational Enterprise: Comment. *American Economic Review* 93:980–994.

Blonigen, B. A., Davies, R. B., Waddell, G. R. and Naughton, H. (2007). FDI in Space: Spatial Autoregressive Relationships in Foreign Direct Investment. *European Economic Review*, 51(5):1303–1325.

Blonigen, B. A. and Feenstra, R. C. (1997). Protection Threats and Foreign Direct Investment. In Feenstra, R.C. (ed.), *The Effects of U.S. Trade Protection and Promotion Policies.* Chicago: National Bureau of Economic Research and University of Chicago Press, pp 55–80.

Blonigen, B. A. and Prusa, T. J. (2003). Antidumping. In Choi, E. K. and Harrigan, J. (eds), *Handbook of International Trade.* Cambridge, MA: Blackwell Publishers, pp 251–284.

Brainard, S. L. (1997). An Empirical Assessment of the Proximity-Concentration Trade-Off Between Multinational Sales and Trade. *American Economic Review*, 87(4):520–544.

Brandao De Brito, J. and De Mello Sampayo, F. (2005). The Timing and Probability of FDI: An Application to the United States Multinational Enterprises. *Applied Economics*, 37(4):417–437.

Brandao de Brito, J. and Mello Sampayo, F. (2004). The Locational Determinants of the U.S. Multinationals Activities. Working Papers, No. 11–04, Banco de Portugal.

Braunerhjelm, P., Ekholm, K., Grundberg, L., and Karpaty, P. (1996). *Swedish Multinational Corporations: Recent Trends in Foreign Activities*. Research Institute of Industrial Economics (IUI), Working Paper No. 462.

Brenton, P., Di Mauro, F. and Lücke, M. (1999). Economic Integration and FDI: An Empirical Analysis of Foreign Investment in the EU and in Central and Eastern Europe. *Empirica*, 26(2):95–121.

Campa, J. M. (1993). Entry by Foreign Firms in the United States Under Exchange Rate Uncertainty. *Review of Economics and Statistics*, 75(4):614–622.

Carr, D. L., Markusen, J. R. and Maskus, K. E. (2001). Estimating the Knowledge-Capital Model of the Multinational Enterprise. *American Economic Review*, 91(3):693–708.

Carstensen, K. and Toubal, F. (2004). Foreign Direct Investment in Central and Eastern European Countries: A Dynamic Panel Data Analysis. *Journal of Comparative Economics*, 32:3–22.

Clausing, K. A. (2000). Does Multinational Activity Displace Trade? *Economic Inquiry*, 38(2):190–205.

De Santis, R. A., Anderton, R. and Hijzen, A. (2004). *On the Determinants of Euro Area FDI to the United States: The Knowledge-Capital-Tobin's Q Framework*. Working Papers Series, No. 329, European Central Bank.

Dixit, A. (1989). Hysteresis, Import Penetration and Exchange Rate Pass-Through. *Quarterly Journal of Economics*, 104(2):205–228.

Egger, P. (2001). European Exports and Outward Foreign Direct Investment. *Weltwirtschaftliches Archiv*, 137(3):427–449.

Egger, P. and Pfaffermayr, M. (2003). The Proper Panel Econometric Specification of the Gravity Equation: A Three-Way Model with Bilateral Interaction. *Empirical Economics,* 28:571–580.

Egger, P. and Pfaffermayr, M. (2004a). Distance, Trade and FDI: A Hausman-Taylor SUR Approach. *Journal of Applied Econometrics*, 19(2):227–246.

Egger, P. and Pfaffermayr, M. (2004b). Foreign Direct Investment and European Integration in the 1990s. *The World Economy* 27(1):99–110.

Falzoni, A. M. (2000). Statistics of Foreign Investment and Multinational Corporations: A Survey. University of Bergamo, Centro de studi Luca d'Agliano and CESPRI.

Feenstra, R. C. (1997). U.S. Exports, 1972–1994: *With State Exports and Other U.S. Data*. NBER Working Paper 5990. Cambridge, MA: National Bureau of Economic Research, Inc.

Feinberg, S. E. and Keane, M. P. (2001). U.S.–Canada Trade Liberalization and MNC Production. *Review of Economics and Statistics* , 83(1):118–132.

Feinberg, S. E. and Keane, M. P. (2006). Accounting for the Growth of MNC-Based Trade Using a Structural Model of U.S. MNCs. *American Economic Review*, 96(5):1515–1558.

Froot, K. A. and Stein J. C. (1991). Exchange Rates and Foreign Diect Investment: An Imperfect Market Approach. *Quarterly Journal of Economics*, 106(4):1191–1217.

Girma, S. (2002). The Process of European Integration and the Determinants of Entry by Non-EU Multinationals in the UK Manufacturing. *The Manchester School*, 70:315–335.

Girma, S., Greenaway, D. and Wakelin, K. (2002). Does Antidumping Stimulate FDI? Evidence from Japanese Firms in the UK. *Review of World Economics*, 3:414–436.

Greene, W. H. (2004). The Behaviour of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects. *Econometrics Journal*, 7(1):98–119.

Grossman, G. and Helpman, E. (1996). Foreign Investment with Endogenous Protection. In R.C. Feenstra, G.M. Grossman, and D.A. Irwin (eds.), *The Political Economy of Trade Policy*, Cambridge, MA: MIT Press.

Guo, J. Q. and Trivedi, P. K. (2002). Firm-Specific Assets and the Link between Exchange Rates and Japanese Foreign Direct Investment in the United States: A Re-Examination. *Japanese Economic Review*, 53(3):337–349.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method Moments Estimators. *Econometrica*, 50:1029–1054.

Hanson, G. H., Mataloni, R. J. and Slaughter M. J. (2005). Vertical Production Networks in Multinational Firms. *Review of Economics and Statistics*, 87(4):664–678.

Hausman, J., Hall, B. H. and Griliches, Z. (1984). Econometric Models for Count Data with an Application to the Patents-R&D Relationship. *Econometrica*, 52(4):909–938.

Head, K. and Ries J. (2001). Overseas Investment and Firm Exports. *Review of International Economics*, 9(1):108–122.

Helpman, E. (1984). A Simple Theory of International Trade with Multinational Corporations. *Journal of Political Economy*, 92(3):451–471.

Helpman, E., Melitz, M. J. and Yeaple, S. R. (2004). Export versus FDI with Heterogeneous Firms. *American Economic Review,* 94(1):300–316.

Hsiao, F. S. T. and Hsiao, M. C. W. (2004). The Chaotic Attractor of Foreign Direct Investment–Why China? A Panel Data Analysis. *Journal of Asian Economics*, 15(4):641–670.

IMF and OECD (2001). *Foreign Direct Investment Statistics: How Countries Measure FDI*. Washington, DC: IMF and OECD.

Klein, M. W., Peek, J. and Rosengren E. S. (2002). Troubled Banks, Impaired Foreign Direct Investment: The Role of Relative Access to Credit. *American Economic Review*, 92(3):664–682.

Klein, M. W. and Rosengren, E. S. (1994). The Real Exchange Rate and Foreign Direct Investment in the United States: Relative Wealth vs. Relative Wage Effects. *Journal of International Economics*, 36(3–4):373–389.

Lipsey, R. (2001) *Foreign Direct Investment and the Operations of Multinational Firms: Concepts, History and Data*. NBER Working Paper 8665. Cambridge, MA: National Bareau of Economic Research, Inc.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge, MA: Cambridge University Press.

Markusen, J. R. (1984). Multinationals, Multi-Plant Economies, and the Gains from Trade. *Journal of International Economics,* 16:205–226.

Markusen, J. R. (1995). The Boundaries of Multinational Enterprise and the Theory of International Trade. *Journal of Economic Perspectives,* 9:169–189.

Markusen, J. R. (1997). *Trade versus Investment Liberalization*. NBER Working Paper 6231. Cambridge, MA: National Bureau of Economic Research, Inc.

Markusen, J. R. (2002). *Multinational Firms and the Theory of International Trade.* Cambridge, MA: MIT Press.

Markusen, J. R. and Maskus, K. E. (2002-a). Discriminating Among Alternative Theories of the Multinational Enterprise. *Review of International Economics,* 10:694–707.

Markusen, J. R. and Maskus, K. E. (2002-b). A Unified Approach to Intra-Industry Trade and Direct Foreign Investment. In Lloyd, P.J. and Lee, H.-H., (eds), *Frontiers of Research in Intra-Industry Trade.* New York: Palgrave Macmillan, pp 199–219.

Markusen J. R. and Venables, A. (2000). The Theory of Endowment, Intra-Industry and Multi-National Trade. *Journal of International Economics,* 52:209–234.

Markusen J. R., Venables, A. , Eby-Konan, D. and Zhang, K. (1996). A Unified Treatment of Horizontal Direct Investment, Vertical Direct Investment, and the Pattern of Trade in Goods

and Services. NBER Working Paper 5696. Cambridge, MA: National Bureau of Economic Research, Inc.

Mátyás, L. (2001). Modelling Export Flows in the APEC Region: Static and Dynamic Gravity Model Approaches. *The Asian Pacific Journal of Economics and Business*, 5(1):97–118.

Mold, A. (2003). The Impact of the Single Market Programme on the Locational Determinants of US Manufacturing Affiliates: An Econometric Analysis. *Journal of Common Market Studies,* 41(1):37–62.

Nicoletti, N., Golub, S., Hajkova, D., Mirza, D., and Yoo, K. Y. (2003). Policies and International Integration: Influences on Trade and Foreign Direct Investment, OECD Working Paper No. 359. OECD Economics Department.

OECD (2000). OECD Economic Outlook, 68. Paris

Phillips, P. and Moon, H. (1999). Linear Regression Theory for Non-Stationary Panel Data. *Econometrica,* 67:1057–1111.

Ray, J. (1991). Foreign Takeovers and New Investments in the United States. *Contemporary Policy Issues*, 9(2):59–71.

Stephan, M. and Pfaffmann, E. (1998). Detecting the Pitfalls of Data on Foreign Direct Investment: A Guide to the Scope and Limits of FDI-Data as an Indicator of Business Activities of Transnational Corporations. Discussion-Paper 98–02, University of Stuttgart.

Swenson, D. L. (2004). Foreign Investment and the Mediation of Trade Flows. *Review of International Economics*, 12(4):609–629.

UNCTAD (2002). World Investment Report, Transnational Corporations and Export Competitiveness. United Nations, Geneva.

Wooldridge, J. M. (2003). Cluster-Sample Methods in Applied Econometrics. *American Economic Review,* 93(2):133–138.

Yeaple, S. R. (2003). The Complex Integration Strategies of Multinationals and Cross Country Dependencies in the Structure of Foreign Direct Investment. *Review of Economics and Statistics*, 85(3):726–734.

Yeyati, A. L., Stein, E. and Daude, C. (2003). Regional Integration and the Location of FDI. Working paper No. 492, Washington, DC: Inter-American Development Bank.

Chapter 21
Stochastic Frontier Analysis and Efficiency Estimation

Christopher Cornwell and Peter Schmidt

Theoretically, a production function gives the maximum possible output with a given set of inputs. This is different from its common regression counterpart, which specifies the conditional mean of output. The production function defines a boundary or "frontier", deviations from which can be interpreted as inefficiency. The econometrics of stochastic frontier analysis (SFA) provides techniques for modelling the frontier concept within a regression framework so that inefficiency can be estimated.

Obviously, the notion of a frontier can be extended to other representations of technology. Further, with behavioral assumptions like cost minimization, allocative inefficiency can be distinguished from the technical errors. We discuss ways to make this distinction empirically, but in this chapter we concentrate primarily on the estimation of production frontiers and measures of technical inefficiency relative to them.

The literature on SFA is now roughly 30 years old and surveys have appeared periodically (Førsund, Lovell and Schmidt (1980), Schmidt (1985–86), Lovell and Schmidt (1988), Bauer (1990) and Greene (1993)). In addition, the literature has been given a textbook treatment by Kumbhakar and Lovell (2000). Aside from reviewing recent advances in SFA, this chapter differs from the earlier surveys in its focus on the use of panel data and attention to questions of econometric and statistical detail.

In general, the frontier specifications we consider are variants of the general panel-data regression model:

$$y_{it} = \alpha_t + \mathbf{x}_{it}'\beta + v_{it} - u_{it} = \alpha_{it} + \mathbf{x}_{it}'\beta + v_{it} \ , \tag{21.1}$$

Christopher Cornwell
Department of Economics, University of Georgia, Athens, GA 30602, USA,
e-mail: cornwl@terry.uga.edu

Peter Schmidt
Department of Economics, Michigan State University, East Lansing, MI 48824, USA,
e-mail: schmidtp@msu.edu

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

where y_{it} is output for firm i $(i = 1,\ldots,N)$ at time t $(t = 1,\ldots,T)$, $\mathbf{x}_{it}$ is a vector of inputs and v_{it} is a random error. In contrast to v_{it}, u_{it} is a one-sided error $(u_{it} \geq 0)$, capturing the shortfall of y_{it} from the frontier, $(\alpha_t + \mathbf{x}'_{it}\beta + v_{it})$, The term "stochastic frontier" follows from the fact that the frontier specification includes v_{it}.

Defining $\alpha_{it} = \alpha_t - u_{it}$, we have a model in which inefficiency is reflected in differences between firms in the intercepts. Various special cases arise depending on the restrictions placed on the α_{it}. The early literature on SFA developed in a pure cross-section $(T = 1)$ context, where identification requires strong assumptions about the distributions of v_i and u_i. The application and extension of panel-data econometrics to SFA grew out dissatisfaction with these assumptions. The first panel frontiers treated inefficiency as a time-invariant firm effect, $\alpha_i = \alpha - u_i$. Estimates of the α_i can be obtained using standard panel techniques and converted into estimates of inefficiency. The time-invariance restriction can substitute for the distributional assumptions necessary for cross-section SFA. Later work on panel frontiers introduced specifications for the α_{it} that relax the time-invariance assumption, while retaining the advantages of panel data.

21.1 Measurement of Firm Efficiency

In general, when we say that a firm produces efficiently, we mean this in both a technical and allocative sense. Here our emphasis will be on technical efficiency, but we will pay some attention to allocative efficiency as well, in both cases following the canonical approach to the measurement problem developed by Farrell (1957).

A firm is technically efficient if it uses the minimal level of inputs given output and the input mix or produces the maximal level of output given inputs. The first definition is formalized in Farrell's *input-based* measure,

$$\mathscr{I}(y, \mathbf{x}) = \min[b : f(b\mathbf{x}) \geq y] \,, \tag{21.2}$$

where $\mathscr{I}$ indicates the proportion of $\mathbf{x}$ necessary to produce y, holding the input ratios constant, and f is a standard, neoclassical (frontier) production function. This measure is illustrated in Fig. 21.1, which depicts an inefficient firm producing output y^A with input vector $\mathbf{x}^A$. Technically efficient production occurs along the isoquant, $\text{Isoq}[L(y^A)] = [\mathbf{x} : \mathscr{I}(y^A, \mathbf{x}) = 1]$, where $L(y) = [x : (y, \mathbf{x})$ is feasible] is the input requirements set. Because only $b\mathbf{x}^A$ is required to produce y^A, both inputs must be scaled back by the factor $(1 - b)$ to achieve technical efficiency.

While this measure is used widely, its appeal diminishes when the input set is not strictly convex (the isoquant is not everywhere downward sloping). For example, the input vector $\mathbf{x}^B$ is technically efficient according to the Farrell input measure, although the same level of output could be produced with less of x_1. In this case, a distinction exists between the isoquant and the efficient subset, $\text{ES}[L(y^A)] = [\mathbf{x} : \mathbf{x} \in L(y^A),$ and $\tilde{\mathbf{x}} \leq \mathbf{x}$ implies $\tilde{\mathbf{x}} \notin L(y^A)]$, with $\text{ES}[L(y^A)] \subseteq \text{Isoq}[L(y^A)]$. In most

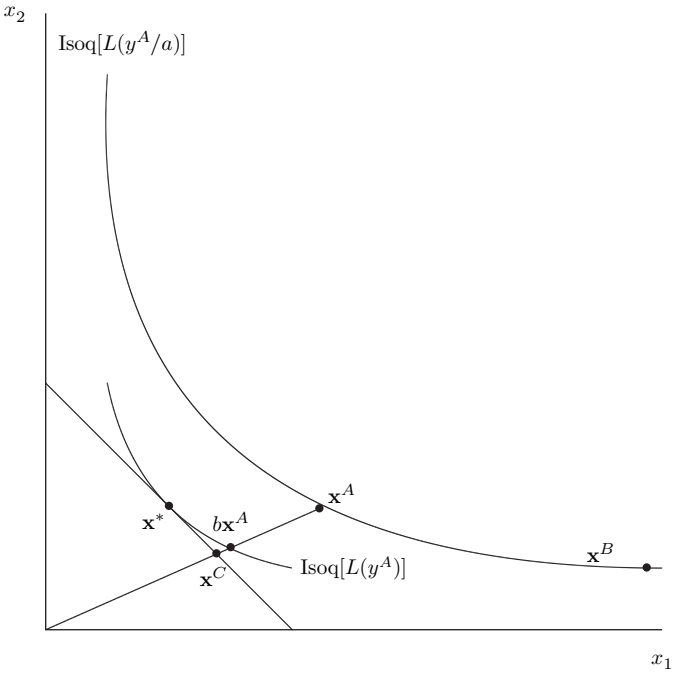

Fig. 21.1 Farrell measures of technical efficiency

econometric specifications this distinction has no practical significance, because the functional forms used in empirical work impose equivalence between the efficient subset and the isoquant (Lovell (1993) and Greene (1993)).

Corresponding to the output-oriented definition of efficiency is Farrell's *output-based* measure,

$$\mathscr{O}(y, \mathbf{x}) = \min\left[a : f(\mathbf{x}) \geq \frac{y}{a}\right]. \tag{21.3}$$

Holding inputs constant, $1/\mathscr{O}$ gives the amount by which output could be expanded. From the perspective of the output-based measure, the firm producing y^A with $\mathbf{x}^A$ in Fig. 21.1 will also be technically efficient if it operates on Isoq$[L(y^A/a)]$.

Färe and Lovell (1978) showed that if f is homogeneous of degree r ($r =$ returns to scale), then $y = f(b\mathbf{x}) = b^r f(\mathbf{x}) = af(\mathbf{x})$ and $a = b^r$. Thus, $\mathscr{I} = \mathscr{O}$ only under constant returns. When technology is not homogeneous, there is no straightforward interpretation of $\mathscr{O}$ in terms of $\mathscr{I}$, a result that has some implications for how technical efficiency is estimated (Atkinson and Cornwell (1994a)).

A firm is allocatively inefficient when the marginal rate of substitution between any two of its inputs is not equal to the corresponding input price ratio. This is true of the firm using $\mathbf{x}^A$ in Fig 21.1, instead of the cost-minimizing input vector $\mathbf{x}^*$. Let $\mathbf{p}$ be the input price vector corresponding to the isocost line through $\mathbf{x}^*$. Then the

(input-based) technical efficiency of the firm producing with $\mathbf{x}^A$ is $b = \mathbf{p}'(b\mathbf{x}^A)/\mathbf{p}'\mathbf{x}^A$, and since $\mathbf{p}'\mathbf{x}^* = \mathbf{p}'\mathbf{x}^C$, its allocative efficiency is the ratio $\mathbf{p}'\mathbf{x}^C/\mathbf{p}'(b\mathbf{x})$. It follows that total or cost efficiency of the firm is given by $\mathbf{p}'\mathbf{x}^C/\mathbf{p}'\mathbf{x}^A$, or the product of technical and allocative efficiency.

21.2 Introduction to SFA

21.2.1 The Basic SFA Empirical Framework

We begin with the Farrell output-based technical efficiency measure in (21.3), which relates observed output, y_i, to the production frontier, $f(\mathbf{x}_i;\beta)$, as follows:

$$y_i = a_i f(\mathbf{x}_i;\beta), \quad 0 < a_i \leq 1\ , \tag{21.4}$$

The basic empirical framework for SFA is a regression specification involving a logarithmic transformation of (21.4) that adds a random error term (v_i), as in

$$\ln y_i = \ln f(\mathbf{x}_i;\beta) + v_i - u_i\ , \tag{21.5}$$

where $u_i = -\ln a_i \geq 0$ represents technical *in*efficiency and output is bounded from above by the *stochastic* frontier $f(\mathbf{x}_i;\beta)\exp(v_i)$. The output-based measure of technical efficiency is obviously recovered as $\exp(-u_i)$.

Models like (21.5) were first introduced by Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977). These papers expressed the view that the frontier specification should be like any other regression function, which is to say, stochastic. Thus, the v_i serve the same purpose as any conventional regression disturbance—to account for random unobserved factors.

The central econometric issue in models like (21.5) is how to treat the u_i. With cross-section data they are usually assumed to follow some non-negative distribution, conditional on $\mathbf{x}_i$. Panel data afford the opportunity to view (21.5) as a standard unobserved-effects model and avoid the distributional assumption. Other issues, such as choosing a functional form and the specification for $f(\mathbf{x}_i;\beta)$, are also important insofar as they affect the estimation of firm efficiency.

21.2.2 Stochastic vs Deterministic Frontiers

The earliest attempts to quantify production inefficiency treated the frontier as deterministic, ignoring the role of v_i. The classic example of this approach is Aigner and Chu (1968). Aigner and Chu calculated β as the solution to either the linear or quadratic programming problem, taking $f(\mathbf{x}_i;\beta)$ to be Cobb-Douglas, and

computed technical inefficiency as deviations from the fitted frontier. By ignoring v_i, all deviations from the frontier were regarded as inefficiency. Further, because there is no stochastic structure to these models, it does not make sense to talk about the statistical properties of their approach.

Closely related to the Aigner–Chu procedure is the non-parametric programming technique of data envelopment analysis (DEA). With DEA the goal is to "envelop" the data with a quasi-convex hull. Since DEA is non-parametric, it is robust to misspecification of the functional form for $f(\mathbf{x}_i;\beta)$. See Cooper, Seiford and Zhu (2004) for a recent survey of DEA.

The analysis of a deterministic frontiers can be made statistical by treating $u_i \equiv y_i - f(\mathbf{x}_i;\beta)$ as random variables. A simple strategy is assume the u_i are iid with a constant mean μ and constant variance, and uncorrelated with $\mathbf{x}_i$. In the Cobb-Douglas setup of Aigner and Chu, this recasts in problem as a regression of the form

$$\ln y_i = \alpha^* + \sum_{k=1}^{K} \beta_k \ln x_{ik} - u_i^* \,, \tag{21.6}$$

where $\alpha^* = (\alpha - \mu)$ and $u_i^* = u_i - \mu$. Ordinary least squares (OLS) consistently estimates α^* and the β_ks, from which a "corrected" OLS (COLS) estimator of α can be obtained:

$$\hat{\alpha} = \hat{\alpha}^* + \max_i(-\hat{u}_i^*) \,, \tag{21.7}$$

where $\hat{u}_i^* = \ln y_i - \hat{\alpha}^* - \sum_k \hat{\beta}_k \ln x_{ik}$. Then, letting $\hat{u}_i$ denote the corrected residuals based on $\hat{\alpha}$, technical efficiencies can be estimated as $\exp(-\hat{u}_i)$. However, the distribution of $\hat{\alpha}$ is unknown even asymptotically.

Likelihood-based approaches to (21.6) exist as well; for example, the solutions to the Aigner–Chu linear (quadratic) programming problem is a maximum-likelihood estimator (MLE) if the u_i are exponential (half-normal) (Schmidt (1976)). Still, the properties of these estimators remain unknown, because the range of y_i depends on β, violating one of the regularity conditions for the usual properties of MLEs to hold.

Similarly, a statistical analysis of DEA is possible if assumptions are made about the nature of the randomness in the data. One possibility that has been suggested is to assume simply that the data points $(y_i, \mathbf{x}_i)$ are a random sample from the set of feasible production points. Under this assumption, plus some regularity conditions on the distribution of these points in the neighborhood of the frontier, the DEA measure is a consistent estimator of the efficiency level of a given firm, and its rate of convergence is known. The asymptotic distribution theory is rather complicated. Bootstrapping is also possible, although there are some non-standard features of the bootstrap that are necessary in this setting. For a survey, see Simar and Wilson (2000).

We do not recommend deterministic frontiers. This is partly due to our philosophical view of the nature of randomness in the world, and partly due to the relative complexity of statistical inference in deterministic frontier models.

21.2.3 Other Frontier Functions

Circumstances and objectives sometimes arise that make alternative representations of technology a more desirable framework for efficiency analysis. These include the presence of multiple outputs, exogeneity assumptions and interest in estimating allocative efficiency.

Recently, it has become popular to accommodate multiple outputs through the use of distance functions (e.g. Coelli and Perelman (1996), Morrison, Johnston and Frengley (2000), and Atkinson, Cornwell and Honerkamp (2003)), which are directly related to the Farrell measures of technical inefficiency. For example, the input distance function is defined as the maximum scale factor necessary to place $\mathbf{x}$ on the boundary of $L(\mathbf{y})$:

$$D_{\mathscr{I}}(\mathbf{y},\mathbf{x}) \equiv \max_{\lambda}\left[\lambda \,:\, \left(\frac{\mathbf{x}}{\lambda}\right) \in L(\mathbf{y})\right], \tag{21.8}$$

where $\mathbf{y}$ is a vector of outputs. The reciprocal of $D_{\mathscr{I}}$ is just the Farrell input measure, which implies b in (21.2) is $1/\lambda$. The empirical counterpart to (21.8) can be expressed as

$$0 = \ln D_{\mathscr{I}}(y_i,\mathbf{x}_i) + v_i - u_i\,, \tag{21.9}$$

where $u_i = -\ln b_i$. Estimation of the technology parameters can proceed as a straightforward application of the generalized method of moments (GMM), since standard assumptions about $\mathbf{x}_i$, v_i and u_i imply a set of moment conditions that identify the model (see Atkinson, Cornwell and Honerkamp (2003)). Extracting estimates of the u_i is possible with the methods described in the next section.

The most commonly adopted strategy for estimating technical and allocative efficiency together is to adopt a cost function framework. One might also use a cost function to accommodate multiple outputs or because it is more reasonable to assume output is exogenous. The usual cost frontier specification is derived from (21.2) as

$$C = g\left(\mathbf{y},\frac{\mathbf{p}}{b}\right) = \min_{b\mathbf{x}}\left[\left(\frac{\mathbf{p}}{b}\right)'(b\mathbf{x}) \,:\, f(b\mathbf{x}) = \mathbf{y}\right] = \frac{1}{b}\,g(\mathbf{y},\mathbf{p})\,, \tag{21.10}$$

where C is observed cost, $\mathbf{p}$ is a vector of input prices and the last equality follows from the fact that a cost function is linearly homogeneous in $\mathbf{p}$. Equation (21.10) leads to empirical models of the form

$$\ln C_i = \ln g(y_i,\mathbf{p}_i) + v_i + u_i\,, \tag{21.11}$$

where $u_i = -\ln b_i$. The u_i in (21.11) measure *cost* efficiency, which will generally include both technical and allocative distortions. Below we discuss how to distinguish between the two sources of error.

21.2.4 SFA with Cross-Section Data

21.2.4.1 Estimating the Basic Stochastic Production Frontier

Estimation of (21.5) usually hinges on distributional assumptions for the v_i and u_i. Such assumptions are required to estimate technical efficiency at the firm level with cross-section data. The usual choices are $v_i \,|\, \mathbf{x}_i \sim N(0, \sigma_v^2)$ and $u_i \,|\, \mathbf{x}_i \sim N^+(0, \sigma_u^2)$ (half-normal). Other possibilities for u_i include exponential, truncated normal and gamma, and evidence suggests that frontier estimates are not robust to the choice (Schmidt and Lin (1984)). Given distributions for v_i and u_i and a functional form selected for $f(\mathbf{x}_i; \beta)$, the standard approach is to estimate (21.5) by ML and is automated in popular econometric software such as Stata, TSP and Limdep. There is also a COLS option for the stochastic frontier case in which the OLS estimator of the intercept is corrected by a consistent estimator of $E(u_i)$, identified through the higher-order moments of the OLS residuals.

In most cases, the whole point of the frontier estimation exercise is to compare efficiencies at the firm level. Thus the focus of estimation ultimately is on the residuals, but no matter how they are computed, they represent $\widehat{(v_i - u_i)}$, not $\hat{u}_i$. Estimation of firm-specific efficiencies requires that $\hat{u}_i$ be extracted from $\widehat{(v_i - u_i)}$.

Jondrow, Lovell, Materov and Schmidt (1982) proposed an estimator for the $\hat{u}_i$ based on $E[u_i \,|\, (v_i - u_i)]$ evaluated at $\widehat{(v_i - u_i)}$. Under the usual assumptions of the model, consistent estimates of the technology parameters can be obtained via ML or OLS, from which the $\widehat{(v_i - u_i)}$ can be calculated. Although the Jondrow et al. estimator is not consistent (because the variation associated with the distribution of u_i conditional on $(v_i - u_i)$ is independent of N), there is no alternative consistent estimator of firm-level efficiency when using cross-section data.

21.2.4.2 Estimating Technical and Allocative Efficiency

Schmidt and Lovell (1979) first demonstrated how to incorporate allocative distortions by introducing errors in the first-order conditions for cost minimization. With distributional assumptions for the allocative errors, they estimated the first-order conditions along with the production frontier. Because Schmidt and Lovell adopted the self-dual Cobb-Douglas functional form, their decomposition of cost efficiency into technical and allocative components was straightforward.

A more typical framework for estimating technical and allocative efficiency jointly is a cost system with the general form,

$$C_i = g(y_i, \mathbf{p}_i) \exp(v_i + u_i + \eta_i) \tag{21.12}$$

$$s_{ik} = s(y_i, \mathbf{p}_i) \exp(\omega_{ik}), \tag{21.13}$$

where s_{ik} is the observed share of the kth input, $s(\cdot)$ is the optimal share implied by Shephard's lemma applied to the deterministic component of the cost function, $g(y_i, \mathbf{p}_i)$, and η_i and ω_{ik} are random disturbances reflecting allocative inefficiency.

There is an inherent econometric challenge in estimating (21.12) and (21.13) with cross-section data, because of the relationship between the allocative errors. Allocative inefficiency raises costs, so η_i must be one-sided, but allocative distortions involve over- and under-utilization of inputs, so the ω_{ik} will be two-sided. Further, η_i and ω_{ik} will be correlated with each other. Without relying on functional form restrictions, or assuming the problem away by asserting independence between η_i and ω_{ik}, estimation is complicated. Kumbhakar (1997) derived a general solution to the problem, but his model is highly nonlinear in the terms representing allocative efficiency and therefore difficult to estimate. More optimistically, Atkinson and Cornwell (1994b) show how panel data can obviate the problem entirely.

21.3 SFA with Panel Data

21.3.1 Models with Time-Invariant Inefficiency

The models we consider in this section are special cases of (21.1), with $\alpha_{it} = \alpha_i = \alpha - u_i$, so that

$$y_{it} = \alpha_i + \mathbf{x}'_{it}\beta + v_{it} \; . \tag{21.14}$$

From the viewpoint of the panel-data literature, (21.14) is just a standard unobserved-effects model. Unless otherwise noted, we maintain the following assumptions for (21.14):

$$(A.1) \quad E(v_{it} \,|\, \mathbf{x}_i^o, \alpha_i) = 0, \; t = 1, \ldots, T \tag{21.15}$$

$$(A.2) \quad E(\mathbf{v}_i \mathbf{v}'_i \,|\, \mathbf{x}_i^o, \alpha_i) = \sigma_v^2 I_T \tag{21.16}$$

where $\mathbf{x}_i^o = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ and $\mathbf{v}_i$ is $T \times 1$. Thus we generally treat the variables in $\mathbf{x}_{it}$ as strictly exogenous (which, in a production context, could perhaps be defended using the argument of Zellner, Kmenta and Dreze (1966)) and require the v_{it} to be conditionally homoscedastic and serially uncorrelated. Approaches to estimating (21.14) differ depending on what is assumed about the α_i (u_i).

From this point on, we will no longer make an explicit notational distinction between a variable and its logarithm. To be consistent with most empirical specifications, we will assume y and $\mathbf{x}$ are measured in logs. Thus, (21.14) can be thought of as a Cobb-Douglas production frontier. However, the form of $f(\mathbf{x}_i; \beta)$ is not very important for how we proceed with estimation, as long as the unobserved effect/inefficiency is additive.

21.3.1.1 Advantages of Panel Data

The use of panel data to fit production relationships dates back at least to Mundlak (1961), who used repeated observations on farms to control for unobserved soil quality and managerial ability that affect output and may be correlated with inputs. In most applications outside the SFA literature, this is the primary motivation for using panel data—to control for unobservables that may be correlated with $\mathbf{x}_{it}$.

The first use of panel data in SFA was by Pitt and Lee (1981), but not until Schmidt and Sickles (1984) was the link between the frontier and panel-data literatures systematically established. They identified three advantages of panel data for SFA. First, the assumption of independence between $\mathbf{x}_i$ and u_i invoked in cross-section estimation can be relaxed. Second, specific distributional assumptions for v_i and u_i, required in cross-section data to estimate efficiency at the firm level, can be avoided. Third, firm-level efficiency can be estimated more precisely, and, in the case where $T \to \infty$, consistently. There is one caveat, however. These benefits come at the expense of another assumption—that inefficiency does not vary over time. The longer the panel, the less sense this assumption makes.

21.3.1.2 Estimating the Basic Panel Frontier Model

It is common in the panel-data literature to say that estimation of (21.14) depends on whether the α_i are fixed or random. As argued originally by Mundlak (1978) and emphasized by Wooldridge (2002), this terminology misses the point. Of course the α_i are random; the issue is whether they are correlated with $\mathbf{x}_{it}$. To take a fixed-effects (FE) approach to estimation is to allow arbitrary correlation between $\mathbf{x}_{it}$ and α_i. A random-effects (RE) specification generally denies this possibility, or allows such correlation only in very specific ways. This point is especially important in the SFA literature, where correlation between inputs and inefficiency (u_i) is a concern.

To facilitate the review of estimator choices for (21.14), we rewrite the model combining all T observations for a single firm:

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{e}_T\alpha_i + \mathbf{v}_i \,, \tag{21.17}$$

where $\mathbf{y}_i$ and $\mathbf{v}_i$ are vectors of length T, $\mathbf{X}_i$ is $T \times K$ and $\mathbf{e}_T$ is a $T \times 1$ vector of ones. We begin the review maintaining assumptions (A.1) and (A.2) and leaving open the possibility that $\mathbf{x}_{it}$ is correlated with u_i. Under these two assumptions, the asymptotically efficient procedure is the FE estimator,

$$\hat{\beta}_{\text{FE}} = \left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{M}_i\mathbf{X}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{M}_i\mathbf{y}_i \,, \tag{21.18}$$

where $\mathbf{M}_i = \mathbf{I}_T - \mathbf{e}_T(\mathbf{e}_T'\mathbf{e}_T)^{-1}\mathbf{e}_T'$ is the familiar projection that transforms the data into deviations from firm means (for example, $y_{it} - \bar{y}_i$, $\bar{y}_i = T^{-1}\sum_t y_{it}$). The estimator is easily computed as OLS of $y_{it} - \bar{y}_i$ on $\mathbf{x}_{it} - \bar{\mathbf{x}}_i$.

More asymptotically efficient estimators exist if correlation between $\mathbf{x}_{it}$ and u_i can be ruled out. An assumption like

$$(A.3) \quad E(u_i \,|\, \mathbf{x}_i^o) = \mu \;, \tag{21.19}$$

does this, where $\mu > 0$ indicates that the u_i are drawn from a one-sided distribution. Notationally, we accommodate μ as in (21.6), by defining $\alpha^* = (\alpha - \mu)$ and $u_i^* = u_i - \mu$. Then, (21.17) becomes

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{e}_T\alpha^* + \varepsilon_i \;, \tag{21.20}$$

where $\varepsilon_i = v_{it} - u_i^*$. Along with $(A.3)$ it is also common to assume

$$(A.4) \quad E(u_i^2 \,|\, \mathbf{x}_i^o) = \sigma_u^2 \;, \tag{21.21}$$

which implies $E(\varepsilon_i\varepsilon_i') \equiv \Omega_i = \sigma_v^2\mathbf{I}_T + \sigma_u^2\mathbf{e}_T\mathbf{e}_T'$. Under $(A.1)$–$(A.4)$, the standard RE estimator,

$$\begin{pmatrix} \hat{\alpha}^* \\ \hat{\beta} \end{pmatrix}_{\text{RE}} = \left[\sum_{i=1}^N (\mathbf{e}_T, \mathbf{X}_i)'\Omega_i^{-1}(\mathbf{e}_T, \mathbf{X}_i)\right]^{-1} \sum_{i=1}^N (\mathbf{e}_T, \mathbf{X}_i)'\Omega_i^{-1}\mathbf{y}_i \;, \tag{21.22}$$

is asymptotically efficient.

Calculating (21.22) is equivalent OLS of $\Omega_i^{-1/2}\, Y_i$ on $\Omega_i^{-1/2}$, where $\Omega_i^{-1/2} = \mathbf{I}_T - (1-\psi)\mathbf{P}_i$, $\psi = [\sigma_v^2/(\sigma_v^2 + T\sigma_u^2)]^{1/2}$ and $\mathbf{P}_i = \mathbf{I}_T - \mathbf{M}_i$. The form of $\Omega_i^{-1/2}$ implies a "quasi-demeaning" of the data, (for example, $y_{it} - (1-\psi)\bar{y}_i$), that subsumes the FE transformation. Clearly, as $T \to \infty$, $\psi \to 0$ and $\hat{\beta}_{\text{RE}} \to \hat{\beta}_{\text{FE}}$. Actual implementation requires consistent estimators for σ_v^2 and σ_u^2. There are a number of alternatives, but the most popular follows Wallace and Hussain (1969) and estimates the variance components using the FE and "between" residuals, which are obtained from OLS of $\bar{y}_i$ on $\bar{\mathbf{x}}_i$.

Occasionally, the RE estimator is justified on the grounds that some of the variables of interest do not vary over time and such variables are swept away by the FE transformation. This is not necessary because the coefficients of time-invariant variables (say $\mathbf{z}_i$) can be estimated as OLS of $(\bar{y}_i - \bar{\mathbf{x}}_i\hat{\beta}_{\text{FE}})$ on $\mathbf{z}_i$. However, the estimated coefficients of $\mathbf{z}_i$ will be consistent only if the time-invariant variables are uncorrelated with u_i. In this case, one would not use the RE estimator either, for the same reason.

Hausman and Taylor (1981) offered a solution to this problem in the form of an efficient instrumental-variables (IV) estimator that allows some variables in $\mathbf{x}_{it}$ and $\mathbf{z}_i$ to be correlated with the u_i. Letting $\mathbf{X}_i^* = (\mathbf{X}_i, \mathbf{Z}_i)$, their estimator can be written as

$$\begin{pmatrix} \hat{\alpha}^* \\ \hat{\beta} \end{pmatrix}_{\text{HT}} = \left[\sum_{i=1}^N (\mathbf{e}_T, \mathbf{X}_i^*)'\Omega_i^{-1/2}\mathbf{P}_{A_i}\Omega_i^{-1/2}(\mathbf{e}_T, \mathbf{X}_i^*)\right]^{-1} \sum_{i=1}^N (\mathbf{e}_T, \mathbf{X}_i^*)'\Omega_i^{-1/2}\mathbf{P}_{A_i}\Omega_i^{-1/2}\mathbf{y}_i \;, \tag{21.23}$$

where $\mathbf{P}_{A_i}$ is the projection onto the instrument set $\mathbf{A}_i = (\mathbf{M}_i\mathbf{X}_i, \mathbf{P}_i\mathbf{X}_{i1}, \mathbf{Z}_{i1})$ and $(\mathbf{X}_{i1}, \mathbf{Z}_{i1})$ designates variables that are uncorrelated with u_i. Identification requires that there are at least as many variables in $\mathbf{X}_{i1}$ as in $\mathbf{Z}_{i2}$. Assuming $(A.1)$, $(A.2)$, $(A.4)$ (appropriately modified to include $\mathbf{z}_i$) hold, along with the Hausman-Taylor orthogonality conditions, $\hat{\beta}_{HT}$ is the efficient GMM estimator using the instruments in $\mathbf{A}_i$.

Amemiya and MaCurdy (1986) show that more instruments are implied by the assumption that $\mathbf{X}_{i1}$ and $\mathbf{Z}_{i1}$ are uncorrelated with u_i. Their efficient IV estimator has the same form as (21.23), but uses the instrument set $[\mathbf{M}_i\mathbf{X}_i, \mathbf{I}_T \otimes (\mathbf{x}_{i1}^o, \mathbf{z}_{i1})]$. Breusch, Mizon and Schmidt (1989) further extend the Hausman-Taylor estimator under the additional assumption that the correlation between $\mathbf{X}_{i2}$ and u_i is constant over time. Regardless of the instrument employed, any estimator of the form given in (21.23) can be carried out by applying the RE data transformation to (21.20) and estimating the transformed regression by IV.

The assumptions that lead to more efficient estimators than $\hat{\beta}_{FE}$ can be tested using the well known methodology of Hausman (1978). A Hausman test of the difference between $\hat{\beta}_{\text{FE}}$ and $\hat{\beta}_{\text{RE}}$ will provide evidence on whether the data support $(A.3)$. The restrictions embodied in the efficient IV estimators can be tested in a similar fashion, or by using the GMM-based test of overidentification suggested by Hansen (1982).

Finally, it is worth pointing out that most popular econometric software automates $\hat{\beta}_{FE}$ and $\hat{\beta}_{\text{RE}}$, and the Hausman test of their difference. Some (for example, Stata and Limdep) also contain procedures to compute the Hausman–Taylor and Amemiya–McCurdy estimators. However, as we have discussed, all of these estimators are easily implemented with standard OLS or IV packages after appropriately transforming the data.

21.3.1.3 Firm-Specific Technical Efficiency Estimates

Given any consistent estimator β, firm-specific estimates of technical inefficiency can be obtained using a COLS procedure as with a deterministic frontier. This involves calculating

$$\hat{u}_i = \hat{\alpha} - \hat{\alpha}_i, \quad \hat{\alpha} = \max_i(\hat{\alpha}_i)\,, \tag{21.24}$$

normalizing the frontier in terms of the best firm in the sample. Then, the remaining firms' efficiency levels are estimated by $\exp(-\hat{u}_i)$, which is consistent as $T \to \infty$ (assuming $\hat{\beta}$ is).

In the FE case, α_i can be estimated as $\hat{\alpha}_i = \bar{y}_i - \bar{\mathbf{x}}_i'\hat{\beta}_{FE}$, or by direct OLS estimation of (21.17) in which the α_i appear as coefficients of firm-specific dummy variables. The latter is cumbersome if the sample contains a large number of firms, but some software packages (Stata and Limdep) offer this as an alternative to their regular FE procedure (OLS on demeaned data). Because the FE estimator of β is consistent under relatively weak conditions, it is appealing as a basis for SFA. However, its appeal diminishes if the empirical frontier specification includes time-invariant

regressors. As we suggested earlier, the effects of these variables can be parsed out of $\hat{\alpha}_i$, but their estimated coefficients will be consistent only if the time-invariant variables are uncorrelated with u_i, and if that is the case the Hausman-Taylor estimator is preferred.

We should point out that, while $\hat{\alpha}$ is consistent as $T \to \infty$, it is biased *upward* when T is fixed. This upward bias is due to the "max" operation, and is consequently more severe the larger N is. In fact, Park and Simar (1994) show that consistency of $\hat{\alpha}$ requires the condition that $(\ln N)/T^{1/2} \to 0$, so that N cannot increase too fast as T increases. The upward bias of $\hat{\alpha}$ in the fixed-T case causes a downward bias in estimated efficiencies (that is, in the $\hat{u}_i$). We will comment more on this issue when we discuss inference on the inefficiencies.

A more difficult problem is distinguishing inefficiency from *unobservable* time-invariant variables. The COLS procedure will overstate a firm's inefficiency if there are time-invariant unobservables, but the alternatives require more assumptions. One example is Heshmati and Kumbhakar (1994), who deal with capital as a fixed unobservable in a study of Swedish dairy farms. Their strategy is to assume that $(y_{it} - \mathbf{x}'_{it}\beta)$ can be decomposed as $(\alpha_i + v_{it} + u_{it})$, with $u_{it} \leq 0$, and treat the α_i as a fixed firm effect (representing unobserved capital). In addition, they take the v_{it} and u_{it} to be conditionally normal and half-normal, as in standard cross-section SFA. They impose independence between $\mathbf{x}_{it}$ and u_{it}, and they also assume that the u_{it} are independent over time. The latter is a very unrealistic assumption. This approach will likely understate inefficiency because any time-invariant component of inefficiency is eliminated with the fixed effects, and any persistent component will be at least partially eliminated.

The RE specification accommodates time-invariant regressors, but care should be taken in testing the assumptions that serve as a basis for estimation, whether the estimator is RE or efficient IV. Residuals constructed from either estimator can be used to estimate $\alpha_i = \alpha^* - u_i^*$ and carry out the COLS procedure in (21.24). If the RE estimator is justified, an alternative is to use the best linear unbiased predictor (BLUP) of u_i^*,

$$\hat{u}_i^* = \frac{-\hat{\sigma}_u^2 \sum_t \hat{\varepsilon}_{it}}{T\hat{\sigma}_u^2 + \hat{\sigma}_v^2} , \tag{21.25}$$

in the COLS procedure.

Finally, as the Heshmati and Kumbhakar (1994) example illustrates, it is also possible to proceed with essentially the same assumptions as in the cross-section case. The only advantage of panel data then is the added precision that comes from repeated observations on each firm. Battese and Coelli (1988) typify this approach, assuming the u_i are truncated normal and the v_{it} are normal, conditional on $\mathbf{x}_i^o$. They estimate α, β and the parameters of the error distributions by ML. An advantage of this approach, if $\mathbf{x}_{it}$ and u_i are independent, is that the frontier intercept α is estimated directly, without the need for the "max" operation in (21.24). Thus, the estimated frontier is not normalized in terms of the best firm and the best firm need not be defined as 100 percent efficient. Battese and Coelli showed how to obtain firm-specific efficiency estimates by generalizing the Jondrow et al. (1982)

decomposition for a panel-data setting. The Battese–Coelli procedure is usually available in software packages that support ML estimation of the cross-section frontier model.

Regardless of the method, the estimation of firm-specific technical inefficiency is straightforward. However, inference regarding the $\hat{u}_i$ is not. This is especially true for the COLS procedure; because of the "max" operation in (21.24), standard distributional results do not apply. We take up the problem of inference in a separate section below.

21.3.1.4 Explaining Firm Efficiency

Often one is interested not only in estimating efficiency levels, but also in determining whether observable firm characteristics can explain them. For example, one might ask whether state-owned or privately owned enterprises differ in their efficiency levels, or whether big firms are more efficient than small firms. Questions like these can be addressed in the context of a stochastic frontier model in which the distribution of technical inefficiency depends on such firm characteristics.

To be more explicit, we consider a stochastic frontier model like (21.1) above, and now assume that the technical inefficiency term u_{it} depends on some observed variables $\mathbf{z}_{it}$, with the dependence expressed as $u_{it}(\mathbf{z}_{it}, \delta)$. We treat these variables as exogenous, so they can include inputs or functions of inputs, but they should not be a function of output.

As a specific example, the model of Reifschneider and Stevenson (1991), Caudill and Ford (1993) and Caudill, Ford and Gropper (1995) (hereafter, RSCFG) assumes that u_{it} is distributed as $N(0, \sigma_{it})^+$, where σ_{it} is a function of $\mathbf{z}_{it}$ and δ. One possibility is $\sigma_{it} = \exp(\mathbf{z}'_{it}\delta)$. Since the expected value of u_{it} is proportional to σ_{it}, we have parameterized the mean of technical inefficiency. However, since the variance of u_{it} is proportional to σ_{it}^2, we have also parameterized its variance. As a result specifications of this type are also referred to as models of heteroskedasticity. Kumbhakar and Lovell (2000) discuss models of heteroskedasticity in one place (Sect. 3.4) and incorporating exogenous influences on efficiency in another (Chap. 7), but in our view these are the same. We will discuss these models from the point of view of explaining efficiency.

Many empirical analyses have proceeded in two steps. In the first step, one estimates the stochastic frontier model and firms' efficiency levels, ignoring $\mathbf{z}$. In the second step, one tries to see how efficiency levels vary with $\mathbf{z}$, perhaps by regressing a measure of efficiency on $\mathbf{z}$. It has long been recognized that such a two-step procedure will give biased results. Since $E(\mathbf{y}|\mathbf{x}, \mathbf{z})$ depends on both $\mathbf{x}$ and $\mathbf{z}$, the first-step regression of $\mathbf{y}$ on $\mathbf{x}$ will be biased by the omission of $\mathbf{z}$, if $\mathbf{x}$ and $\mathbf{z}$ are correlated. A more subtle point is that the calculation of the firm-specific inefficiencies depends on the variances of v_{it} and u_{it}. Ignoring the fact that the variance of u_{it} is not constant, these estimates will be under-dispersed. These points are discussed in Kumbhakar and Lovell (2000), page 119 and Chap. 7, and in Wang and

Schmidt (2002), Sect. 2.3. Simulations reported in Wang and Schmidt indicate these biases are very severe. Accordingly, we do not recommend two-step procedures.

The alternative to a two-step procedure is one-step estimation by ML. For example, under the assumptions of the RSCFG model, the density for observation i,t is well defined. The likelihood then follows from an additional assumption on the independence, or form of dependence, over time at the firm level. This will be discussed in more detail below.

The literature contains several alternatives to the RSCFG setup described above. One is the model employed by Kumbhakar, Ghosh and McGuckin (1991), Huang and Liu (1994), and Battese and Coelli (1995) (hereafter, KGMHLBC), which assumes that the distribution of u_{it} is $N(\mu_{it},\sigma_{it})^+$. So, compared to the RSCFG specification, this model parameterizes the mean rather than the variance of the pre-truncation normal distribution. Several possibilities have been suggested for the parameterization of μ_{it}, including $\mu_{it} = \mathbf{z}'_{it}\delta$ and $\mu_{it} = \mu \cdot \exp(\mathbf{z}'_{it}\delta)$. The KGMHLBC model is heavily used in empirical applications, in part because it is readily available in the FRONTIER software (Coelli 1996). Another is the model of Wang (2002), in which the distribution of u_{it} is $N(\mu_{it},\sigma^2_{it})^+$, and where μ_{it} and σ_{it} both depend on $\mathbf{z}_{it}$. Wang's model allows for non-monotonic effects of $\mathbf{z}_{it}$ on u_{it} and can be used to test the adequacy of the simpler specifications.

We now return to the point made above about the nature of dependence over time. The simplest assumption, and the one most commonly made, is that (conditional on $\mathbf{z}_{i1},\ldots,\mathbf{z}_{iT}$) the u_{it} are independent over time. Since the v_{it} are also typically assumed to be independent over time, the errors $(v_{it}-u_{it})$ are independent over time, and the likelihood is just the product, over all i and t, of the density for observation i,t. It is widely recognized that the independence assumption is unrealistic. It is less widely recognized that the MLE assuming independence is consistent even if the independence assumption is false. In this case, however, a non-standard (robust) covariance matrix calculation is required for the estimates. This is a textbook point in the more general panel-data context, and is discussed in the frontiers context by Alvarez, Amsler, Orea and Schmidt (2004).

Some of the models reviewed in this section satisfy the scaling property that $u_{it} = h(\mathbf{z}_{it},\delta)\cdot\tilde{u}_{it}$, where $\tilde{u}_{it}$ does not depend on $\mathbf{z}_{it}$. For example, the RSCFG model has this property, with $\tilde{u}_{it}$ distributed as $N(0,1)^+$, and with the scaling function $h(\mathbf{z}_{it},\delta)$ equal to the parameterized function σ_{it}. If the scaling property holds, Battese and Coelli (1992) show how to construct the likelihood under the assumption that the underlying random variable $\tilde{u}_{it}$ is time-invariant (and hence just equals $\tilde{u}_i$). However, no model currently exists that allows correlation over time in a less restricted form.

21.3.1.5 Inference Based on Estimated Efficiencies

One of the advantages of SFA (over approaches based on deterministic frontier specifications) is the ability to measure the uncertainty of efficiency estimates. In addition to providing point estimates of a firm's level of efficiency, confidence intervals and hypothesis tests can be constructed.

First, consider an error-components setup with the v_{it} assumed to iid normal and the u_i are iid truncated normal, conditional on $\mathbf{x}_i^o$, as in Battese and Coelli (1988). They show that the u_i conditional on $(v_{i1}-u_i, v_{i2}-u_i, \ldots, v_{iT}-u_i)$ have a normal distribution truncated from below at zero. The mean and variance (before truncation) of the normal distribution are given by (21.9) and (21.10) of Battese and Coelli; the mean depends on the average residual for the firm. The suggested point estimate (or prediction) for u_i is the mean of the truncated distribution, as given by their (21.11). However, we can also obtain confidence intervals for u_i directly from this distribution; for example, a 95 percent confidence interval for u_i is given by the range between the 2.5 and 97.5 percentiles of the truncated normal conditional distribution of u_i. This possibility was first noted by Horrace and Schmidt (1996). Similar methods apply for inefficiency defined as $\exp(-u_i)$. Similar comments also apply in the cross-sectional case when the method of Jondrow et al. (1982) is used to estimate u_i.

Matters are more complicated under weaker assumptions that prevail in most panel settings. Recall the COLS estimator for u_i given in (21.24). Standard results give the joint distribution of the $\hat{\alpha}_i$, and the difficult nature of the inferential problem is due to the max operation. To emphasize this point, for the moment we will ignore the possible inaccuracy of the max operation in picking the maximal population intercept. Suppose that the maximal estimated intercept is $\hat{\alpha}_m$, where m represents a specific observation, and note that α_m may or may not be the maximal intercept in the population. Then, $\hat{u}_i = \hat{\alpha}_m - \hat{\alpha}_i$ and we can use standard methods to construct a confidence interval for $\alpha_m - \alpha_i$. For example, if the v_{it} are normal or if T is large, confidence intervals would be based on the Student's t or standard normal distributions.

There is also an extensive literature on *multiple* comparison procedures. A good general discussion is given by Hochberg and Tamhane (1987). These procedures allow the construction of simultaneous confidence intervals for the $(N-1)$-dimensional vector of differences $(\alpha_m - \alpha_i,\ i \neq m)$. This is a "multiple comparison with a control" (MCC) problem, since for the moment we are treating α_m as a control, or standard of comparison, without being concerned about whether it is in fact the maximal population intercept. Dunnett (1955) gives an easily computable solution to the MCC problem for the special case that the α_j are equicorrelated, and relevant tabulations are given in Hochberg and Tamhane (1987), Dunnett (1964), Dunn and Massey (1965) and Hahn and Hendrickson (1971). Horrace and Schmidt (2000) provide evidence that the equicorrelated assumption is very nearly met in some applications and discuss approximate solutions when it is not met. These confidence intervals may encompass both positive and negative values because they do not assume that α_m is the maximal population intercept.

From the SFA perspective, we are interested in simultaneous confidence intervals for the N-dimensional vector of differences $(\alpha - \alpha_i,\ i = 1, \ldots, N)$, where α is the maximal intercept in the population rather than in the sample. This is a "multiple comparison with the best" (MCB) problem which differs from the MCC problem because it is not assumed that we know which observation corresponds to the maximal population intercept. This problem was solved by Edwards and Hsu (1983),

who showed how MCB intervals could be constructed from MCC intervals. Other relevant references include Hsu (1981, 1984), and a survey is given by Horrace and Schmidt (2000). The MCB intervals give non-negative lower and upper bounds for the differences $u_i = \alpha - \alpha_i$, and the lower bound equals zero for a subset of the firms. The MCB intervals are wider than the corresponding MCC intervals because they include uncertainty about which observation is best. Some empirical examples of MCB intervals are given in Sect. 21.4.

Another possible method of inference based on the FE estimates is bootstrapping. We will begin with a very brief discussion of bootstrapping in the general setting where we have a parameter θ, and there is an estimator $\hat{\theta}$ based on a random sample $(z_1, \ldots, z_N)$. The following bootstrap procedure will be repeated many times, say for $b = 1, \ldots, B$ where B is large. For iteration b, construct "pseudo data", $z_1^{(b)}, \ldots, z_N^{(b)}$, by sampling randomly with replacement from the original data. From the pseudo data, construct the estimate $\hat{\theta}^{(b)}$. The basic result of the bootstrap is that under fairly general conditions the asymptotic (large-N) distribution of $(\hat{\theta}^{(b)} - \hat{\theta})$ conditional on the sample is the same as the (unconditional) asymptotic distribution of $(\hat{\theta} - \theta)$. Thus, for large N the distribution of $\hat{\theta}$ around θ is the same as the bootstrap distribution of $\hat{\theta}^{(b)}$ around $\hat{\theta}$, which is revealed by the large number of bootstrap draws.

We now consider the application of the bootstrap to the specific case of the FE estimates. Our discussion follows Simar (1992). Define the residuals based on the FE estimates of β and α_i as $\hat{v}_{it} = y_{it} - \hat{\alpha}_i - \mathbf{x}_{it}\hat{\beta}_{FE}$. The bootstrap samples will be drawn by resampling these residuals, because the v_{it} are the quantities analogous to the z_is in the previous paragraph, in the sense that they are assumed to be iid, and they are the observable versions of the v_{it}. (The sample size N above corresponds to NT.) So, for bootstrap iteration $b = 1, \ldots, B$, we calculate the bootstrap sample $\hat{v}_{it}^{(b)}$ and the pseudo data, $y_{it} = \hat{\alpha}_i + \mathbf{x}_{it}\hat{\beta}_{FE} + \hat{v}_{it}^{(b)}$. From these data we get the bootstrap estimates of the inefficiencies, and the bootstrap distribution of these estimates is used to make inferences about the actual inefficiencies.

We note that the estimates depend on the quantity $\max_i \hat{\alpha}_j$. Since "max" is not a smooth function, it is not immediately apparent that this quantity is asymptotically normal, and if it were not the validity of the bootstrap would be in doubt. A rigorous proof of the validity of the bootstrap for this problem is given by Hall, Härdle and Simar (1995). They prove the equivalence of the following three statements: (i) $\max_i \hat{\alpha}_j$ is asymptotically normal; (ii) the bootstrap is valid as $T \to \infty$ with N fixed; and (iii) there are no ties for $\max_i \hat{\alpha}_i$, that is, there is a unique index i such that $\alpha_i = \max_i \alpha_j$. There are two important implications of this result. First, the bootstrap will not be reliable unless T is large. Second, this is especially true if there are near ties for $\max_i \alpha_j$, in other words, when there is substantial uncertainty about which firm is best.

Simulation results reported in Kim (1999) are fairly pessimistic. The bootstrap does not lead to very reliable inference on the individual firm efficiencies unless T is very large, or the variance of v_{it} is quite small.

A final possibility for inference is to be a Bayesian. In a Bayesian analysis one postulates a prior distribution for the parameters of the problem, and combines the prior with the likelihood to obtain a posterior distribution upon which inference is based. In models like those we consider here, and in fact in many other models, this inference is done by drawing from the posterior distribution using Markov-Chain Monte Carlo methods.

We will begin with a "Bayesian FE model", due to Koop, Osiewalski and Steel (1997). They postulate an "uninformative" prior for the parameters β, σ_v^2 and α_i. If the v_{it} are iid normal, the mean of the posterior distribution of β is the usual FE estimate, which explains the name of the model. Now consider the inefficiency terms $u_i^* = \max_j \alpha_j - \alpha_i$ or the inefficiencies $\exp(-u_i^*)$. An important point is that an uninformative (flat) prior for the α_i implies a flat prior for the u_i^*, but a (very) informative prior for $\exp(-u_i^*)$. In fact, the prior for $\exp(-u_i^*)$ is proportional to $[\exp(-u_i^*)]^{-1}$, which very, very strongly favors low efficiencies. In a sense this is the Bayesian counterpart to the downward bias of the efficiency estimates using FE that was discussed in Sect. 21.3.1.3. Indeed, the empirical results given in Kim and Schmidt (2000) show a strong similarity between inferences based on the Bayesian FE results and inferences based on bootstrapping the FE estimates.

Koop, Osiewalski and Steel also discuss RE Bayesian models, in which a proper, informative prior is used for the u_i (not the u_i^*). In this model, we estimate absolute rather than relative efficiency, and we treat β, σ_v^2, the overall intercept α and the inefficiencies u_i or $\exp(-u_i)$ as parameters. They consider, for example, independent exponential priors for the u_i. Kim and Schmidt find, unsurprisingly, that the results from a Bayesian analysis with exponential prior inefficiency are quite similar to the results from classical MLE if an exponential distribution is assumed for inefficiency, and the Battese–Coelli result is used to extract the efficiencies. If such results are generally true, as they probably are, it suggests that it does not make much difference whether one is a Bayesian or not; it just matters how strong the assumptions are that one is willing to make about the efficiency distribution. An interesting point is that in this case it is probably easier to be a Bayesian, in a numerical sense, and it also allows more flexibility in choice of distribution.

21.3.1.6 Estimating Technical and Allocative Efficiency

Recall the cost frontier and share equations given in (21.12) and (21.13), where allocative inefficiency is reflected in a one-sided disturbance in the cost equation (η_i) and a two-sided error in the share equations (ω_{ik}), while technical inefficiency is represented solely through a one-sided cost equation error (u_i). As we noted in Sect. 21.2.4.2, the choices for estimating such a system of equations are to either adopt a restrictive functional form, assume η_i and ω_{ik} are independent, or attempt to estimate the specification proposed by Kumbhakar (1997). Although the latter captures the salient features of the relationship between the η_i and ω_{ik} and does not impose a functional-form restrictions, the specification is difficult to estimate because it is highly nonlinear in the terms involving the allocative errors.

An alternative to using error components is to model deviations from cost-minimizing behavior in terms of parameters that scale prices. In this case, the firm is assumed to minimize *shadow cost*, recognizing that although the input mix may be incorrect when judged in terms of market prices, it can be seen as efficient when related to *shadow prices*. The firm minimizes actual costs (is allocatively efficient) only if the ratio of shadow prices equals the ratio of market prices. This parametric approach was developed in a cross-section context by Lau and Yotopoulos (1971), and later extended by Toda (1976), Lovell and Sickles (1983) and Atkinson and Halvorsen (1984).

Atkinson and Cornwell (1994b) generalized the parametric approach to a panel data setting. Reformulating (21.10) as a shadow cost-minimization problem, they consider the estimation of a system of equations like

$$C_{it}^{*} = g^{*}(y_{it}, \mathbf{p}_{it}^{*}) \exp(v_{it} + u_i) \tag{21.26}$$

$$s_{itk}^{*} = s^{*}(y_{it}, \mathbf{p}_{it}^{*}) \exp(\omega_{itk}), \tag{21.27}$$

where $p_{itk}^{*} = \phi_{ik} p_{itk}$ is a vector of shadow prices where the ϕ_{ik} are parameters to be estimated. Because allocative inefficiency is identified through the ϕ_{ik}, the difficulty of fully specifying the relationship between cost and share-equation allocative errors is obviated. Further, the ω_{itk} can be viewed (appropriately) as conventional random errors.

From a panel-data perspective, the system in (21.26) and (21.27) is an unobserved-effects model, where the effects appear as slope coefficients as well as additive intercept terms. As Atkinson and Cornwell show, FE estimation of such a model is straightforward. Firm-specific technical efficiency estimates can be constructed from the $\hat{u}_i$ using COLS. Estimates of ϕ_{ik} indicate the relative over ($\hat{\phi_{ik}} < 1$) or under-utilization ($\hat{\phi_{ik}} > 1$) of an input. Together, the $\hat{u}_i$ and $\hat{\phi_{ik}}$ can be translated into an estimate of the potential cost savings from eliminating inefficiency.

21.3.2 Models with Time-Varying Inefficiency

While there are great benefits to treating efficiency as if it is time-invariant, time invariance is a strong assumption, especially in longer panels. Now we relax this assumption, explicitly taking up the the SFA model given in (21.1), where α_t defines the frontier intercept in period t. In the context of this model, the problem is simple in principle. A firm's level technical efficiency in each time can be estimated period in COLS fashion as $\exp(-\hat{u}_{it})$, where

$$\hat{u}_{it} = \hat{\alpha}_t - \hat{\alpha}_{it}, \quad \hat{\alpha}_t = \max_i(\hat{\alpha}_{it}) \,. \tag{21.28}$$

In practice, however, we cannot expect to identify the α_{it} without placing some additional structure on the frontier model. Different papers have restricted the α_{it} in

different ways. Below we consider the most widely utilized specifications as well as some more recently proposed alternatives.

21.3.2.1 The Model of Cornwell, Schmidt and Sickles

Cornwell, Schmidt and Sickles (1990) (hereafter CSS) approached the problem from the standpoint of a panel regression model with individual-specific slope coefficients:

$$y_{it} = \mathbf{x}_{it}'\beta + \mathbf{w}_{it}'\delta_i + v_{it} \,, \tag{21.29}$$

where $\mathbf{w}_{it}$ is an $L \times 1$ vector of variables whose coefficients, δ_i, vary over i. Clearly, (21.29) is a special case of the production frontier in (21.1) with $\alpha_{it} = \mathbf{w}_{it}'\delta_i$. CSS provide an empirical illustration based on the specification $\mathbf{w}_{it}' = [1, t, t^2]$ so that

$$\alpha_{it} = \mathbf{w}_{it}'\delta_i = \delta_{i1} + \delta_{i2}\, t + \delta_{i3}\, t^2 \,. \tag{21.30}$$

Obviously if $\mathbf{w}_{it}$ contains only a constant, (21.29) reduces to the usual unobserved effects model, which is to say, the basic panel frontier with time-invariant technical efficiency.

In addition to proposing a specification for α_{it}, CSS extend the standard FE and RE panel estimators to models like (21.29). To discuss these estimators, consider the expression of (21.29) that combines all T observations on a single firm:

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{W}_i\delta_i + \mathbf{v}_i \,, \tag{21.31}$$

where $\mathbf{W}_i$ is a $T \times L$ matrix. In addition, assume

$$(A.1') \quad E(v_{it} \,|\, \mathbf{x}_i^o, \mathbf{w}_i^o, \alpha_i) = 0,\ t = 1, \ldots, T \tag{21.32}$$

$$(A.2') \quad E(\mathbf{v}_i\mathbf{v}_i' \,|\, \mathbf{x}_i^o, \mathbf{w}_i^o, \alpha_i) = \sigma_v^2 I_T, \tag{21.33}$$

parallel to (21.15) and (21.16) in Sect. 21.4.1.

As shown by CSS, the extension of the FE estimator in (21.18) is

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{M}_{W_i}\mathbf{X}_i \right)^{-1} \sum_{i=1}^{N} \mathbf{X}_i'\mathbf{M}_{W_i}\mathbf{y}_i \,, \tag{21.34}$$

where $\mathbf{M}_{W_i} = \mathbf{I}_T - \mathbf{W}_i(\mathbf{W}_i'\mathbf{W}_i)^{-1}\mathbf{W}_i'$ is a generalization of the demeaning projection, $\mathbf{M}_i$. Under $(A.1')$ and $(A.2')$, $\hat{\beta}_{FE}$ is consistent and asymptotically normal, but note that identification requires $L \leq T$.

The RE estimator in (21.22) can be likewise extended with the addition of assumptions parallel to (21.19) and (21.21):

$$(A.3') \quad E(\delta_i \,|\, \mathbf{x}_i^o, \mathbf{w}_i^o) = \delta_o \tag{21.35}$$

$$(A.4') \quad E(\delta_i\delta_i' \,|\, \mathbf{x}_i^o, \mathbf{w}_i^o) = \Delta \,. \tag{21.36}$$

After invoking $(A.3')$ and $(A.4')$ and writing δ_i as $\delta_o + \zeta_i$, (21.31) becomes

$$\begin{aligned} \mathbf{y}_i &= \mathbf{X}_i\beta + \mathbf{W}_i\delta_o + \varepsilon_i \\ \varepsilon_i &= \mathbf{W}_i\zeta_i + \mathbf{v}_i. \end{aligned} \tag{21.37}$$

The RE estimator of β and δ_o is given by

$$\begin{pmatrix} \hat{\beta} \\ \hat{\delta}_o \end{pmatrix}_{\text{RE}} = \left[\sum_{i=1}^{N} (\mathbf{X}_i, \mathbf{W}_i)'\Omega_i^{-1}(\mathbf{X}_i, \mathbf{W}_i)\right]^{-1} \sum_{i=1}^{N} (\mathbf{X}_i, \mathbf{W}_i)'\Omega_i^{-1}\mathbf{y}_i \,, \tag{21.38}$$

where now $\Omega_i = \text{cov}(\varepsilon_i) = \sigma_u^2\mathbf{I}_T + \mathbf{W}_i\Delta\mathbf{W}_i'$. Consistent estimators of σ_u^2 and Δ are provided in CSS. Under $(A.1')$–$(A.4')$, $\hat{\beta}_{\text{RE}}$ is asymptotically efficient, but this claim hinges on $(A.3')$.

CSS also extended the Hausman–Taylor efficient IV estimator to the model with individual-specific slope coefficients. This means partially relaxing $(A.3')$ and allowing some of the variables in $(\mathbf{X}_i, \mathbf{W}_i)$ to be correlated with δ_i. Assuming there are enough orthogonality conditions to satisfy identification requirements, CSS show that β and δ_o can be estimated as

$$\begin{aligned} \begin{pmatrix} \hat{\beta} \\ \hat{\delta}_o \end{pmatrix}_{\text{HT}} = &\left[\sum_{i=1}^{N} (\mathbf{X}_i, \mathbf{W}_i)'\Omega_i^{-1/2}\mathbf{P}_{A_i^*}\Omega_i^{-1/2}(\mathbf{X}_i, \mathbf{W}_i)\right]^{-1} \\ &\times \sum_{i=1}^{N} (\mathbf{X}_i, \mathbf{W}_i)'\Omega_i^{-1/2}\mathbf{P}_{A_i^*}\Omega_i^{-1/2}\mathbf{y}_i, \end{aligned} \tag{21.39}$$

where $\mathbf{P}_{A_i^*}$ is the projection onto the transformed instrument set $\mathbf{A}_i^* = \Omega_i^{-1/2}\mathbf{A}_i$ and $\mathbf{A}_i$ is the natural extension of the original Hausman and Taylor instrument set. Although CSS do not pursue it, (21.39) encompasses extensions to the Amemiya–MaCurdy and Breusch–Mizon–Schmidt estimators as well. The estimator in (21.39) is the efficient GMM estimator under assumptions $(A.1')$, $(A.2')$, $(A.4')$ and the orthogonality conditions imposed by $\mathbf{A}_i$. It is worth pointing out that, unlike in (21.23), this efficient-GMM equivalence depends on the use of transformed instruments in $\mathbf{P}_{A_i^*}$. Although the RE data transformation is more complicated in this case, in principle $\hat{\beta}_{HT}$ can be computed by premultiplying (21.38) by $\Omega_i^{-1/2}$ and performing IV using $\mathbf{A}_i^*$ as instruments.

Firm-specific technical inefficiencies can be estimated using methods directly analogous to those Sect. 21.3.1.3. In the FE case, this involves estimating the elements of δ_i either by OLS of $y_{it} - \mathbf{x}_{it}'\hat{\beta}_{FE}$ on $\mathbf{w}_{it}$ or directly as coefficients of firm dummies interacted with $\mathbf{w}_{it}$. Then compute the $\hat{\alpha}_{it}$ as $\mathbf{w}_{it}'\hat{\delta}_i$ and $\hat{u}_{it}$ as in (21.28). Because the frontier intercept may vary from period to period, the temporal pattern of technical efficiency will vary from firm to firm. Consider, for example, the CSS specification of α_{it} given in (21.30). Although $\hat{\alpha}_{it}$ will be quadratic in t for each firm, $\hat{\alpha}_t$ may not be, which implies u_{it} may not be either. The setup for estimating technical efficiencies is essentially the same in the RE case, whether β and δ_o are estimated by RE or or efficient IV. The only difference is the set of residuals used in the calculations.

21.3.2.2 The Models of Kumbhakar and Battese and Coelli

Kumbhakar (1990) and Battese and Coelli (1992) proposed time-varying efficiency specifications of the form

$$u_{it} = \gamma(t, \theta)\delta_i \,, \tag{21.40}$$

where $\delta_i \geq 0$ is a scalar and $\gamma(t, \theta)$ is a scalar function of time and a vector of parameters, θ. Kumbhakar assumes $\gamma(t, \theta) = [1 + \exp(bt + ct^2)]^{-1}$, with $\theta = (b, c)$. Depending on the values of b and c, the temporal pattern of inefficiency could be increasing or decreasing, concave or convex. Battese and Coelli propose an alternative model, $\gamma(t, \theta) = 1 + \eta_1(t - T) + \eta_2(t - T)^2$, where $\theta = (\eta_1, \eta_2)$.

Because $\gamma(t, \theta)$ does not vary by firm in (21.40), the temporal pattern of technical efficiency is the same for all firms, in contrast to CSS. Also different from CSS, Kumbhakar and Battese and Coelli couch their specifications in panel extensions of the classic cross-section SFA model introduced in Sect. 21.2. Thus, estimation of their models depends on distributional assumptions for δ_i and v_{it} that impose independence between efficiency and $\mathbf{x}_{it}$. Kumbhakar and Battese and Coelli derive the MLEs for their respective models (treating the δ_i as truncated normal and v_{it} as normal, conditional on $\mathbf{x}_{it}$) and show how to estimate firm-specific technical efficiencies by extending Jondrow et al. (1982).

It is possible to estimate the models of Kumbhakar and Battese and Coelli under weaker conditions than they imposed. For example, we could assume ($A.4'$) instead, and integrate (21.40) into a RE panel regression model like (21.38) as follows:

$$y_{it} = \mathbf{x}_{it}'\beta - \gamma(t, \theta)\delta_o + [v_{it} - \gamma(t, \theta)(\delta_i - \delta_o)] \,. \tag{21.41}$$

Such model can be estimated by nonlinear least squares and firm-specific technical efficiencies obtained using the procedure in (21.28). All that is required is a simple regression of firm-i residuals on $\gamma(t, \theta)$ to estimate $(\delta_i - \delta_o)$. However, FE estimation of (21.41) is econometrically more complicated because the unobserved effects do not enter additively. This point will be discussed more fully in the next section.

Finally, the connection between specifications like (21.40) and those similar to (21.30) is straightforward when we express the former in terms of α_{it}. Suppose, instead of (21.40), we asserted that $\alpha_{it} = \gamma(t, \theta)\delta_i$. So long as $\gamma(t, \theta)$ is positive for all t, then $\alpha_t = \max_j(\alpha_{jt}) = \gamma(t, \theta)\max_j(\delta_j)$ and $u_{it} = \gamma(t, \theta)[\max_j(\delta_j - \delta_i)]$, so that the α_{it} and u_{it} have the same temporal pattern, determined by the function $\gamma(t, \theta)$, and this pattern is the same for all firms.

21.3.2.3 The Model of Ahn, Lee and Schmidt

The models presented in Sects. 21.4.1 and 21.4.2 allow technical inefficiency to vary over time, but in a structured way. In this section we consider an alternative model that was originally proposed by Kiefer (1980), and which was subsequently applied to the frontiers problem by Lee and Schmidt (1993), and further analyzed by Ahn, Lee and Schmidt (2001) and Han, Orea and Schmidt (2005). In this model the

temporal pattern of inefficiency is arbitrary, but (as in (21.40) above) it is restricted to be the same for all firms. The specification is

$$\alpha_{it} = \lambda_t \delta_i \,, \tag{21.42}$$

where the λ_t are parameters to be estimated. One can think of (21.42) as a special case of (21.40) with $\gamma(t, \theta)$ represented by a set of time dummies. As such, an advantage of (21.42) is that any parametric form such as Kumbhakar's is a testable special case. See, for example, Bai (2003).

The RE estimator of this model raises no new issues, but its FE estimator is interesting. We consider assumptions similar to $(A.1')$ and $(A.2')$; that is, strict exogeneity of the regressors and the white noise property of the errors. Ahn, Lee and Schmidt propose GMM estimators that impose the restrictions implied by these assumptions. An surprising result is that the moment conditions based on the white noise assumption are useful (result in an increase in asymptotic efficiency) even if the errors are normal. This is certainly not the case in the usual linear regression model without fixed effects. They also analyze the true FE estimator, defined by the minimization of $\sum_i \sum_t (y_{it} - \mathbf{x}_{it}'\beta - \lambda_t \delta_i)^2$ with respect to β, λ_t and δ_i. The consistency of this estimator requires the white noise assumption. Also, given the white noise assumption, this estimator has a non-standard form for its covariance matrix, and it is less efficient than the efficient GMM estimator, even if the errors are normal. Once again these are results that are not true in the linear regression model without fixed effects.

Han, Orea and Schmidt (2005) extend this analysis to the case that λ_t is a parametric function of time and some parameters. Therefore they make possible a FE analysis of models like those of Kumbhakar or Battese and Coelli (discussed in the previous section). The essential results of Ahn, Lee and Schmidt extend to this case. This means that a true FE analysis is possible, but it depends on a white noise assumption, and it requires a non-standard calculation of the covariance matrix of the estimates.

21.4 Applications

In this section we will discuss two empirical applications of the techniques that this paper has described. References to additional applications can be found in the survey papers listed in Sect. 21.1 above.

21.4.1 Egyptian Tile Manufacturers

First, we review the analysis of Egyptian tile manufacturers as originally conducted by Seale (1985, 1990). The author personally collected data on a set of firms in the

Fayoum region of Egypt, and was involved in, but did not supervise, the collection of data in another region (Kalyubiya). The total sample consisted of 9 firms in the Fayoum region and 16 in Kalyubiya. Data were collected over a 66-week period in 1982–1983. This time period was divided into 3-week time intervals, so that the maximum number of observations was 22. However, because firms did not produce in all periods, the actual number of observations was not the same for all firms; that is, the panel was unbalanced.

The firms make floor tiles using a rather simple technology: sand, cement and water are mixed and pressed into tiles, which are dried in the sun. Three types of tiles are made, and the firms do not produce other products. The capital of the firms consists of a few types of machines: mixers, electric presses, manual presses and polishers. There are only two skill categories of workers.

The original data were aggregated into measures of output, labor and capital. (Because the physical inputs are used in essentially fixed proportions to output, labor and capital were the only inputs to be included in the production function.) Because of the relatively small number of output and input types, and because the data were collected personally by the individual conducting the analysis, the aggregation process is probably much less troublesome than in the typical production function analysis.

The basic empirical results were generated in 1984 and 1985 and used the methodology available at that time; namely, the MLE of Pitt and Lee (1981) and the FE and RE estimators of Schmidt and Sickles (1984), suitably modified to account for the unbalanced nature of the panel. A Cobb-Douglas production function was assumed. Hausman tests rejected the RE specifications, and so the focus was on the FE treatment. The estimated coefficient of capital (machine hours) was positive but insignificant, while the estimated coefficient of labor was insignificantly different from unity. Thus, for all practical purposes, estimated efficiency differences reflect differences in output per worker-hour.

Firm efficiencies were estimated separately for the two areas, since they were viewed as distinct markets. The estimates of technical efficiency ranged from 100 to 71 percent in the Fayoum area and from 100 to 56 percent in the Kalyubiya area. This is a reasonable range given the costs of transporting output, and the least efficient firms were located in small and remote villages where competition from larger and more efficient firms was not a real threat.

Seale argues convincingly that his efficiency estimates do indeed reflect differences that one might interpret as inefficiency (as opposed to measurement error, omitted inputs, etc.). For example, consider the following description of an inefficient firm (Seale (1985, page 175)): “The organization of the firm could be improved; the working area around the electric press is organized for three workers only, while many tileries with an electric press are able to provide adequate space for four workers to form tiles. The total working area, though large, is cluttered with broken tiles and empty sacks, giving a general impression of disarray.” Furthermore, Seale ranked the firms in terms of their apparent efficiency after his initial visits to them, but before the data were collected and analyzed. His a priori rankings were very similar to those from the statistical analysis. In fact, the rank correlation

coefficient was 0.98 for the Fayoum region, in which he lived and supervised the data collection effort, and 0.72 in Kalyubiya. This is fairly convincing evidence, unfortunately of a type that we cannot expect generally to be available, that the efficiency estimation exercise has been basically successful.

These data have subsequently been analyzed by a number of others, including Horrace and Schmidt (1996, 2000) and Kim and Schmidt (2000). The following results are from Kim and Schmidt. For reasons of space we will quote only the results for one firm, number 4, which is the median-efficient firm based on the FE estimates.

The FE estimates yield an efficiency level for firm 4 of 0.895. A set of 90 percent MCB intervals give a confidence interval for firm 4 of [0.648, 1]. The "one" here is exact—it is not the result of rounding. The usual percentile bootstrap gives a 90 percent confidence interval of [0.692, 0.940]. The Bayesian FE model gives a point estimate (mean of the posterior distribution) of 0.812, which is somewhat lower, and a 90 percent confidence interval (this is not a Bayesian word, but it is a Bayesian calculation) of [0.688, 0.945]. Note the similarity of the Bayesian interval to the interval from bootstrapping the FE estimates.

RE models give results that are relatively similar. For the half-normal MLE, the point estimate of efficiency for firm 4 is 0.885 and a 90 percent confidence interval, based on the Battese–Coelli method, is [0.787, 0.978]. For the exponential MLE, we obtain 0.896 and [0.799, 0.984], and the Bayesian exponential model with an uninformative prior for the exponential parameter yields 0.891 and [0.782, 0.986].

Kim and Schmidt argue that these results are optimistic, in the sense that the choice of specific model is not too important, and the results are precise enough to be of some potential use.

21.4.2 Indonesian Rice Farmers

Next, we turn to the analysis of Indonesian rice farmers. These data have been analyzed by Erwidodo (1990), Lee and Schmidt (1993) and Horrace and Schmidt (1996). The data contain information on 171 rice farms in Indonesia, for six growing seasons. They were collected by the Agro Economic Survey, as part of the Rural Dynamic Study in the rice production area of the Cimanuk River Basin, West Java, and obtained from the Center for Agro Economic Research, Ministry of Agriculture, Indonesia. In particular, they were not collected as primary data by the individuals later involved in the analysis, though Erwidodo was personally familiar with farming practices in the area. Time periods are growing seasons, of which there are two per year; three of the six time periods are dry seasons and three are wet seasons. The data were collected from six different villages that contain 19, 24, 37, 33, 22 and 36 farm families, respectively. This is a balanced panel in the sense that every family is observed for the same six time periods.

Output is production of rough rice, in kilograms. The inputs include seed, urea, tri-sodium phosphate (TSP), labor and land area. Erwidodo considered both

Cobb-Douglas and translog specifications, but we will follow Lee and Schmidt and discuss only results for the Cobb-Douglas specification; this does not make much difference. Besides the inputs, the equation that is estimated also includes some dummy variables, as follows. DP is a dummy variable equal to one if pesticides are used, and zero otherwise. DV1 equals one if high-yielding varieties of rice are planted, while DV2 equals one if mixed varieties are planted; the omitted category represents traditional varieties. DSS equals one in the wet season and zero otherwise. DR1, ... , DR5 are dummy variables representing the six villages, and are intended to control for differences in soil quality or other relevant factors across villages. Finally, DSIZE is a dummy variable equal to one if the land area is greater than 0.5 hectare. Erwidodo included this variable while Lee and Schmidt did not, but in fact it makes little difference to the efficiency estimation exercise. We will report results only for the specification that does not include DSIZE. The data are described in detail in Erwidodo (1990).

Erwidodo estimated the model using the standard panel-data techniques: OLS, FE and RE estimators. The results based on the three methods are quite similar; correspondingly, the appropriate Hausman test failed to reject the RE specification. The estimated coefficients of the five input variables were all positive and significant at the usual critical levels. The elasticities ranged from 0.47 for land area to 0.078 for TSP, using the RE estimates, and from 0.43 to 0.09 using the FE estimates. Returns to scale were insignificantly different from unity. The coefficient estimates of the dummies for rice variety and for wet season were significantly different from zero, while the rest of the dummy variables were usually insignificant. The results that were significant indicate that high-yielding rice varieties have higher yields than traditional varieties, and that output is higher in the wet season than in the dry season.

Erwidodo calculates measures of both technical and allocative inefficiency, but we will discuss measures of technical inefficiency only. He calculates estimates of technical inefficiency in three ways: (i) the simple FE calculation given in (21.24) above; (ii) the RE calculation involving the best linear predictor, given in (21.25) above; and (iii) the method of Battese and Coelli (1988). Thus in Erwidodo's implementation of method (iii), distributional assumptions are used in the separation of inefficiency from noise even though they were not used in estimation. We should also note that we might expect the FE results (i) to differ rather substantially from the RE results (ii) or (iii) since in the FE regression we cannot include the time-invariant village dummy variables, and thus differences across villages in soil quality or other relevant time-invariant factors are not controlled.

Erwidodo actually reports his results only for method (iii). Battese and Coelli assumed that u_i is distributed as $N(\mu, \sigma_u^2)$ truncated below at zero. Erwidodo assumed $\mu = 0$ so that u_i is half-normal. In this case $\text{var}(u_i) = \sigma_u^2(\pi - 2)/\pi$. The usual variance components estimates that are part of the RE procedure yield an estimate of $\text{var}(u_i)$, and this can be converted into an estimate of σ_u^2 by multiplying it by $\pi/(\pi - 2)$. It appears that Erwidodo used his estimate of $\text{var}(u_i)$ as an estimate of σ_u^2, neglecting the factor $\pi/(\pi - 2)$, which made his technical inefficiency figures too small. Horrace and Schmidt (1996) recalculation of Erwidodo's results yields

farm-specific inefficiency estimates ranging from 3.5 to 25.8 percent, with a mean of 10.6 percent.

Using the same data, Lee (1991) calculates technical inefficiency measures based on FE estimation and method (i) above. Technical inefficiency now ranges from zero to 64.6 percent, with a mean of 56.7 percent. Estimation by RE and use of method (ii) gives results that are very similar to those for the FE estimator; for example, mean technical inefficiency is then 57.1 percent. These results are consistent with Erwidodo's report that FE and RE generated much higher levels of technical inefficiency that the Battese–Coelli method, but that all three methods give similar rankings.

Clearly there are striking differences between these results. To interpret them, it is interesting to look at the precision of the estimates, as reflected in the relevant confidence intervals. These results are given in Horrace and Schmidt (1996, 2000) and Kim and Schmidt (2000). As in the previous section, here we report the results only for the median firm, number 15. For this firm, the FE estimates give an efficiency level of 0.554. The 90 percent MCB confidence interval is [0.300, 1] and the percentile bootstrap interval is [0.398, 0.646]. The Bayesian FE model gives a point estimate of 0.509 and a 90 percent interval of [0.383, 0.656]. Once again the Bayesian FE estimates are similar to the classical FE estimates and the bootstrap.

As we saw above, the RE efficiencies are much higher. For the half-normal MLE, the point estimate of efficiency for firm 15 is 0.923 with a 90 percent confidence interval of [0.792, 0.990]. For the exponential MLE, we obtain 0.935 and [0.834, 0.996], and for the Bayesian exponential model with uninformative prior on the exponential parameter we get 0.935 and [0.823, 0.996].

Clearly these results are less precise than for the previous data set, and the choice of technique matters more. Kim and Schmidt argue that this is a difficult data set to analyze, because T is fairly small and because the variance of noise (v) is large relative to the variance of inefficiency (u). In this case we can gain a lot of precision by putting more structure on the model, but unfortunately the choice of what structure to impose influences the results more strongly. There is no obvious solution to this problem other than to analyze data that have more favorable characteristics.

Lee (1991) and Lee and Schmidt (1993) have also applied the time-varying efficiency model of subsection 3.2.3 to the Erwidodo data. Compared to the simpler model with time-invariant efficiency, this model does not make much difference in the estimates of the technical parameters (regression coefficients) or in the average level of inefficiency. It does yield an interesting temporal pattern of inefficiency (see Fig. 8.1, page 251, of Lee and Schmidt), with significantly higher efficiency levels in time periods $t = 3$ and $t = 4$ than in the other time periods. However, given the confidence intervals reported in the previous paragraphs for the simpler model, it might be argued that a model with less structure is the opposite of what is needed for this application.

21.5 Concluding Remarks

In this chapter, we have given given a broad survey of the stochastic frontier approach to efficiency measurement, with an emphasis on the use of panel data. While a considerable number of details were discussed, we have tried to emphasize two main points. The first main point is that it is really a misuse of words to discuss the measurement of efficiency; properly, we should refer to *estimation* of efficiency. The estimation of efficiency is essentially a statistical problem, in the sense that the results are subject to uncertainty, and this is true whether traditional statistical methods are used or not. There are two main advantages to an explicitly statistical approach, such as is possible using stochastic frontier models. First, an accommodation can be made to statistical noise. Second, measures of the uncertainty of the results can be generated. Our empirical results in Sect. 21.6 show the importance of this second point. Using a deterministic (non-statistical) model does not remove this uncertainty; it only hides it.

Our second main point is that panel data are useful because they allow weaker assumptions or greater precision under a given set of assumptions, than would be possible with a single cross section. Most of the work so far on the use of panel data for efficiency estimation has emphasized the possibility of weakened assumptions and more flexible models. In retrospect, this may have been a mistake. Certainly we should suspect that the usual trade-off between flexibility of the model and precision of results applies. If efficiency estimates were more routinely reported along with appropriate measures of the uncertainty associated with them, this trade-off could be made more intelligently.

References

Ahn, S. C., Y. H. Lee and P. Schmidt (2001): GMM Estimation of Linear Panel Data Models with Time-Varying Individual Effects, *Journal of Econometrics*, 101, 219–256.

Aigner, D. J. and S. Chu (1968): On Estimating the Industry Production Function, *American Economic Review*, 58, 826–839.

Aigner, D. J., C. A. K. Lovell and P. Schmidt (1977): Formulation and Estimation of Stochastic Frontier Production Function Models, *Journal of Econometrics*, 6, 21–37.

Alvarez, A., C. Amsler, L. Orea and P. Schmidt (2004): Interpreting and Testing the Scaling Property in Models Where Inefficiency Depends on Firm Characteristics, unpublished manuscript.

Amemiya, T. and T. E. MaCurdy (1986): Instrumental Variable Estimation of an Error-Components Model, *Econometrica*, 54, 869–891.

Atkinson, S. E. and C. Cornwell (1994a): Estimating Output and Input Technical Efficiency Using a Flexible Functional Form and Panel Data, *International Economic Review*, 35, 245–256.

Atkinson, S. E. and C. Cornwell (1994b): Parametric Measurement of Technical and Allocative Inefficiency with Panel Data, *International Economic Review*, 35, 231–244.

Atkinson, S. E. and C. Cornwell and O. Honerkamp (2003): Measuring Productivity Change Using a Malmquist Index: Stochastic Distance Function Estimation vs DEA, *Journal of Business and Economic Statistics*, 21, 284–295.

Atkinson, S. E. and R. Halvorsen (1984): Parametric Efficiency Tests, Economies of Scale, and Input Demand in U.S. Electric Power Generation, *International Economic Review*, 25, 647–662.

Bai, J. (2003): Inferential Theory for Factor Models of Large Dimensions, *Econometrica*, 71, 135–172.

Battese, G. E. and T. J. Coelli (1988): Prediction of Firm-Level Technical Efficiencies with a Generalized Frontier Production Function and Panel Data, *Journal of Econometrics*, 38, 387–399.

Battese, G. E. and T. J. Coelli (1992): Frontier Production Functions, Technical Efficiency and Panel Data: With Applications to Paddy Farmers in India, *Journal of Productivity Analysis*, 3, 153–169.

Battese, G.E. and T. J. Coelli (1995): *A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data, Empirical Economics*, 20, 325–332.

Bauer, P. (1990): Recent Developments in the Econometric Estimation of Frontiers, *Journal of Econometrics*, 46, 39–56.

Breusch, T. S., G. E. Mizon and P. Schmidt (1989): Efficient Estimation Using Panel Data, *Econometrica*, 57, 695–700.

Caudill, S. B. and J. M. Ford (1993): Biases in Frontier Estimation Due to Heteroskedasticity, *Economics Letters*, 41, 17–20.

Caudill, S. B., J. M. Ford and D. M. Gropper (1995): Frontier Estimation and Firm-Specific Inefficiency Measures in the Presence of Heteroskedasticity, *Journal of Business and Economic Statistics*, 13, 105–111.

Coelli, T. J. (1996): A Guide to FRONTIER Version 4.1: A Computer Program for Stochastic Frontier Production and Cost Function Estimation, CEPA Working Paper 96/7, Department of Econometrics, University of New England, Armidale NSW Australia.

Coelli, T. J. and S. Perelman (1996): Efficiency Measurement, Multiple-Output Technologies and Distance Functions: with Application to European Railways, CREPP working paper, 96/05, University of Liège, Belgium.

Cooper, W., L. Seiford and J. Zhu (2004): *Handbook on Data Envelopment Analysis*, Boston: Kluwer Academic Publishers.

Cornwell, C., P. Schmidt and R. C. Sickles (1990): Production Frontiers with Time-Series Variation in Efficiency Levels, *Journal of Econometrics*, 46, 185–200.

Dunn, O. J. and F. J. Massey (1965): Estimation of Multiple Contrasts using t-distribution, *Journal of the American Statistical Association*, 60, 573–583.

Dunnett, C. W. (1955): A Multiple Comparison Procedure for Comparing Several Treatments with a Control, *Journal of the American Statistical Association*, 50, 1096–1121.

Dunnett, C. W. (1964): New Tables for Multiple Comparisons with a Control, *Biometrics*, 20, 482–491.

Edwards, D. G. and J. C. Hsu (1983): Multiple Comparisons with the Best Treatment, *Journal of the American Statistical Association*, 78, 965–971.

Erwidodo (1990): *Panel Data Analysis on Farm-Level Efficiency, Input Demand and Output Supply of Rice Farming in West Java, Indonesia*, unpublished Ph. D. dissertation, Department of Agricultural Economics, Michigan State University.

Färe, R. and C. A. K. Lovell (1978): Measuring the Technical Efficiency of Production, *Journal of Economic Theory*, 19, 150–162.

Farrell, M. S. (1957): The Measurement of Productive Efficiency, *Journal of the Royal Statistical Society*, A, 120, 253–281.

Førsund, F., C. A. K. Lovell and P. Schmidt (1980): A Survey of Frontier Production Functions and of Their Relationship to Efficiency Measurement, *Journal of Econometrics*, 13, 5–25.

Greene, W. H. (1993): The Econometric Approach to Efficiency Analysis, in H. Fried, C. A. K. Lovell and S. Schmidt eds., *The Measurement of Productive Efficiency*, New York: Oxford University Press.

Hahn, G. J. and R. W. Hendrickson (1971): A Table of Percentage Points of the Distribution of the Largest Absolute Value of k Student t Variables and its Applications, *Biometrika*, 58, 323–332.

Hall, P., W. Härdle and L. Simar (1995): Iterated Bootstrap with Applications to Frontier Models, *Journal of Productivity Analysis*, 6, 63–76.

Han, C., L. Orea and P. Schmidt (2005): Estimation of a Panel Data Model with Parametric Temporal Variation in Individual Effects, *Journal of Econometrics*, 126, 241–267.

Hansen, L. (1982): Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica*, 50, 1029–1054

Hausman, J. (1978): Specification Tests in Econometrics, *Econometrica*, 46, 1251–1271.

Hausman, J. and W. Taylor (1981): Panel Data and Unobservable Individual Effects, *Econometrica*, 49, 1377–1399.

Heshmati, A. and S. Kumbhakar (1994): Farm Heterogeneity and Technical Efficiency: Some Results from Swedish Dairy Farms, *Journal of Productivity Analysis*, 5, 45–61.

Hochberg, Y. and A. C. Tamhane (1987): *Multiple Comparison Procedures*, New York: Wiley.

Horrace, W. C. and P. Schmidt (1996): Confidence Statements for Efficiency Estimates from Stochastic Frontier Models, *Journal of Productivity Analysis*, 7, 257–282.

Horrace, W. C. and P. Schmidt (2000): Multiple Comparisons with the Best, with Economic Applications, *Journal of Applied Econometrics*, 15, 1–26.

Hsu, J. C. (1981): Simultaneous Confidence Intervals for All Distances from the Best, *Annals of Statistics*, 9, 1026–1034.

Hsu, J.C. (1984): Constrained Simultaneous Confidence Intervals for Multiple Comparisons with the Best, *Annals of Statistics*, 12, 1145–1150.

Huang, C. J. and J. T. Liu (1994): Estimation o f a Non-Neutral Stochastic Frontier Production Function, *Journal of Productivity Analysis*, 5, 171–180.

Jondrow, J., C. A. K. Lovell, I. Materov and P. Schmidt (1982): On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model, *Journal of Econometrics*, 19, 233–238.

Kiefer, N. M. (1980): A Time Series – Cross Section Model with Fixed Effects with an Intertemporal Factor Structure, unpublished manuscript, Cornell University.

Kim, Y. (1999): *A Study in Estimation and Inference on Firm Efficiency*, unpublished PhD dissertation, Department of Economics, Michigan State University.

Kim, Y. and P. Schmidt (2000): A Review and Empirical Comparison of Bayesian and Classical Approaches to Inference on Efficiency Levels in Stochastic Frontier Models, *Journal of Productivity Analysis*, 14, 91–118.

Koop, G., J. Osiewalski and M. Steel (1997): Bayesian Efficiency Analysis through Individual Effects: Hospital Cost Frontiers, *Journal of Econometrics*, 76, 77–106.

Kumbhakar, S. (1990): Production Frontiers, Panel Data, and Time-Varying Technical Inefficiency, *Journal of Econometrics*, 46, 201–212.

Kumbhakar, S. (1997): Modelling Allocative Efficiency in a Translog Cost Function and Cost Shares: An Exact Relationship, *Journal of Econometrics*, 76, 351–356.

Kumbhakar, S., S. Ghosh and J. T. McGuckin (1991): A Generalized Production Frontier Approach for Estimating Determinants of Inefficiency in U.S. Dairy Farms, *Journal of Business and Economic Statistics*, 9, 279–286.

Kumbhakar, S. and C. A. K. Lovell (2000): *Stochastic Frontier Analysis*, Cambridge: Cambridge University Press.

Lau, L. J. and P. A. Yotopoulos (1971): A Test for Relative Efficiency and an Application to Indian Agriculture, *American Economic Review*, 61, 94–109.

Lee, Y. H. (1991): *Panel Data Models with Multiplicative Individual and Time Effects: Applications to Compensation and Frontier Production Functions*, unpublished Ph. D. dissertation, Department of Economics, Michigan State University.

Lee, Y. H. and P. Schmidt (1993): A Production Frontier Model with Flexible Temporal Variation in Technical Efficiency, in H. Fried, C. A. K. Lovell and S. Schmidt eds., *The Measurement of Productive Efficiency*, New York: Oxford University Press.

Lovell, C. A. K. (1993): Production Frontiers and Productive Efficiency, in H. Fried, C. A. K. Lovell and S. Schmidt eds., *The Measurement of Productive Efficiency*, New York: Oxford University Press.

Lovell, C. A. K. and P. Schmidt (1988): A Comparison of Alternative Approaches to the Measurement of Productive Efficiency, in A. Dogramaci and R. Färe, eds., *Applications of Modern Production Theory: Efficiency and Production*, Boston: Kluwer Academic Publishers.

Lovell, C. A. K. and R. C. Sickles (1983): Testing Efficiency Hypotheses in Joint Production, *Review of Economics and Statistics*, 65, 51–58.

Meeusen, W. and J. van den Broeck (1977): Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error, *International Economic Review*, 18, 435–444.

Morrison, C., W. Johnston and G. Frengley (2000): Efficiency in New Zealand Sheep and Cattle Farming: The Impacts of Regulatory Reform, *Review of Economics and Statistics*, 82, 325–337.

Mundlak, Y. (1961): Empirical Production Function Free of Management Bias, *Journal of Farm Economics*, 43, 44–56.

Mundlak, Y. (1978): On the Pooling of Time-Series and Cross-Section Data, *Econometrica*, 46, 69–85.

Park, B. and L. Simar (1994): Efficient Semiparametric Estimation in a Stochastic Frontier Model, *Journal of the American Statistical Association*, 89, 929–936.

Pitt, M. and L. Lee (1981): The Measurement and Sources of Technical Inefficiency in the Indonesian Weaving Industry, *Journal of Development Economics*, 9, 43–64.

Reifschneider, D. and R. Stevenson (1991): Systematic Departures from the Frontier: A Framework for the Analysis of Firm Inefficiency, *International Economic Review*, 32, 715–723.

Schmidt, P. (1976): On the Statistical Estimation of Parametric Frontier Production Functions, *Review of Economics and Statistics*, 58, 238–239.

Schmidt, P. (1985–86): Frontier Production Functions, *Econometric Reviews*, 4, 289–328.

Schmidt, P. and T. F. Lin (1984): Simple Tests of Alternative Specifications in Stochastic Frontier Models, *Journal of Econometrics*, 24, 349–361.

Schmidt, P. and C. A. K. Lovell (1979): Estimating Technical and Allocative Inefficiency Relative to Stochastic Production and Cost Frontiers, *Journal of Econometrics*, 9, 343–366.

Schmidt, P. and R. C. Sickles (1984): Production Frontiers and Panel Data, *Journal of Business and Economic Statistics*, 2, 367–374.

Seale, J. L. Jr. (1985): *Fixed Effect Cobb-Douglas Production Functions for Floor Tile Firms, Fayoum and Kalyubiya, Egypt, 1981–1983*, unpublished Ph. D. dissertation, Departments of Economics and Agricultural Economics, Michigan State University.

Seale, J. L. Jr. (1990): Estimating Stochastic Frontier Systems with Unbalanced Panel Data: The Case of Floor Tile Manufactories in Egypt, *Journal of Applied Econometrics*, 5, 59–74.

Simar, L. (1992): Estimating Efficiencies from Frontier Models with Panel Data: A Comparison of Parametric, Non-Parametric and Semi-Parametric Methods with Bootstrapping, *Journal of Productivity Analysis*, 3, 171–203.

Simar, L. and P. W. Wilson (2000): Statistical Inference in Nonparametric Frontier Models: The State of the Art, *Journal of Productivity Analysis*, 13, 49–78.

Toda, Y. (1976): Estimation of a Cost Function when Cost is not a Minimum: the Case of Soviet Manufacturing Industries, 1958–1971, *Review of Economics and Statistics*, 58, 259–268.

Wallace T. D. and A. Hussain (1969): The Use of Error Components Models in Combining Cross-Section and Time-Series Data, *Econometrica*, 37, 55–72.

Wooldridge, J. (2002): *Econometric Analysis of Cross-Section and Panel Data*, Cambridge: MIT Press.

Wang, H. J. (2002): Heteroskedasticity and Non-Monotonic Efficiency Effects in a Stochastic Frontier Model, *Journal of Productivity Analysis*, 18, 241–253.

Wang, H. J. and P. Schmidt (2002): One-Step and Two-Step Estimation of the Effects of Exogenous Variables on Technical Efficiency Levels, *Journal of Productivity Analysis*, 18, 129–144.

Zellner, A., J. Kmenta and J. Dreze (1966): Specification and Estimation of Cobb-Douglas Production Functions, *Econometrica*, 34, 784–795.

Chapter 22
Econometric Analyses of Linked Employer–Employee Data

John M. Abowd, Francis Kramarz and Simon Woodcock

22.1 Introduction

There has been a recent explosion in the use of linked employer-employee data to study the labor market. This was documented, in part, in our *Handbook of Labor Economics* chapter (Abowd and Kramarz, 1999a).[1] Various new econometric methods have been developed to address the problems raised by integrating longitudinal employer and employee data. We first described these methods in Abowd and Kramarz (1999b). In this chapter, we present a survey of these new econometric methods, with a particular emphasis on new developments since our earlier articles.

Linked employer-employee data bring together information from both sides of the labor market. They therefore permit, for the first time, equilibrium analyses of labor market outcomes. They also allow researchers to investigate the joint role of worker and firm heterogeneity, both observed and unobserved, on labor market outcomes. Labor economists have taken full advantage of these data to revisit classic questions and to formulate new ones, and much has been learned as a result. For example, Abowd, Kramarz, Lengermann, and Roux (2005) have revisited the classic question of inter-industry wage differentials to determine whether they are

John M. Abowd
School of Industrial and Labor Relations, Cornell University, Ithaca, NY 14850, USA,
e-mail: john.abowd@cornell.edu

Francis Kramarz
CREST-INSEE, Paris, France, e-mail: kramarz@ensae.fr

Simon Woodcock
Department of Economics, Simon Fraser University, Burnaby, BC, Canada,
e-mail: simon_woodcock@sfu.ca

[1] See also Lane, Burgess and Theeuwes (1998) for a review of uses of longitudinal linked employer-employee data.

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

attributable to workers or firms. Abowd, Kramarz, Lengermann and Perez-Duarte (2003) use linked employer-employee data to examine whether "good" workers are employed by "good" firms. Dostie (2005) presents new evidence on the returns to seniority and its relation to turnover; and Woodcock (2003) examines the role of heterogeneity and worker-firm learning on employment and wage dynamics. These applied endeavors have demonstrated the value of linked employer-employee data. They have also spurred the development of new econometric methods to analyze these data. These new methods, rather than specific applications, are the primary focus of this chapter.

A distinguishing feature of longitudinal linked employer–employee data is that individuals and their employers are identified and followed over time. Furthermore, the relation between employer and employee, called a job relation, is continuously monitored. From a statistical perspective, there are three populations under simultaneous study. Individuals are sampled from the population of households, workplaces are sampled from the population of businesses, and jobs are sampled from the population of employment histories. Because of the multiple sampling frames involved, it is necessary to be precise about the statistical structure of the variables under study, since they may come from the individual, employer, or job frame. Measured characteristics of the individual, employer, and job are collected at multiple points in time, which may or may not be synchronous. To make clear the importance of careful elaboration of the sample structure for the variables under study, we will consider a prototypical integrated employer–employee database before turning to specific statistical models. The specific statistical models that we consider are generalizations of the specifications we first used in Abowd, Kramarz and Margolis, (1999, AKM hereafter) as well as in more recent research.

We have noted a general misunderstanding of some recent, and some not-so-recent, empirical methods used by statisticians. We therefore make an effort to relate these methods to those used by panel data econometricians. We show the relation between various fixed effects estimators and estimators popular in the variance components literature – in particular, mixed-effects estimators (see Searle, Casella and McCulloch (1992). As we will see, statisticians and econometricians have different parameters of interest, the former relying more on the variance components and the design of the data, the latter being more concerned with endogeneity in its various guises. These generate a variety of distinct computational issues. Consequently econometricians and statisticians have independently developed a variety of tools to estimate the effects of interest. However, the realized effects have the same interpretation under all methods that we consider.

We begin, in Sect. 22.2, by describing a prototypical longitudinal linked data set and discussing the related problems of missing data and sampling from integrated data. In Sect. 22.3, we present two specifications for linear statistical models that relate linked employer and employee data to outcomes measured at the individual level. In the first and more general specification, person effects and firm effects can reflect interaction between observable person or firm characteristics and unobserved person and firm effects. For instance, match effects are potentially representable in this setting. In the second and simpler specification, a typical individual

has a zero mean for the measured outcomes. Person effects measure deviations over time from this zero mean that do not vary as the employee moves from firm to firm. Firm effects measure deviations from this zero mean that do not vary as the firm employs different individuals. We continue, in Sect. 22.4, by defining a variety of effects that are functions of the basic person and firm effects. Section 22.5 considers the estimation of the person and firm effects by fixed-effects methods. Section 22.6 discusses the use of mixed-effects estimators, the question of orthogonal design, and their relation with various correlated random-effects specifications. In Sect. 22.7 we discuss the important heterogeneity biases that arise when either the person or firm effects are missing or incompletely specified. We discuss the consequences of endogenous mobility in Sect. 22.8, and conclude in Sect. 22.9.

22.2 A Prototypical Longitudinal Linked Data Set [2]

To summarize the complete likelihood function for linked longitudinal employer-employee data, we adopt the formalization in Abowd and Woodcock (2001). They considered statistical models for imputing missing data in linked databases using the full-information techniques developed by Rubin (1987). Their prototypical longitudinal linked data set contains observations about individuals and their employers linked by means of a work history that contains information about the jobs each individual held with each employer. The data are longitudinal because complete work history records exist for each individual during the sample period and because longitudinal data exist for the employer over the same period.

Suppose we have linked data on N workers and J firms with the following file structure. There are three data files. The first file contains data on workers, U, with elements denoted u_i, $i = 1,\ldots,N$. In the discussion below these data are time-invariant but in other applications they need not be. Call U the individual characteristics file. The second data file contains longitudinal data on firms, Z, with elements z_{jt}, $j = 1,\ldots,J$ and $t = 1,\ldots,T_j$. Call Z the employer characteristics file. The third data file contains work histories, W, with elements w_{it}, $i = 1,\ldots,N$ and $t = 1,\ldots,T_i$. Call W the work history file. It contains data elements for each employer who employed individual i during period t. The data U and W are linked by a person identifier. The data Z and W are linked by a firm identifier; we conceptualize this by the link function $j = J(i,t)$ which indicates the firm j at which worker i was employed at date t. For clarity of exposition, we assume throughout that all work histories in W can can be linked to individuals in U and firms in Z and that the employer link $J(i,t)$ is unique for each (i,t).[3]

[2] This section is based on Abowd and Woodcock (2001).

[3] The notation to indicate a one-to-one relation between work histories and indviduals when there are multiple employers is cumbersome. See Abowd and Stinson (2003) for a complete development of the likelihood function allowing for multiple employers during the period.

22.2.1 Missing Data

Abowd and Woodcock consider the problem of imputing missing data in a longitudinal linked database. Their approach is based on the Sequential Regression Multivariate Imputation (SRMI; see Ragunathan et al., 2001). When imputing missing data in each of the three files, they condition the imputation on as much available information as possible. For example, when imputing missing data in the individual characteristics file U they condition not only on the non-missing data in U (observed characteristics of the individual) but also on characteristics of the jobs held by the individual (data in W) and the firms at which the individual was employed (data in Z). Similarly, when conditioning the imputation of missing data in W and Z, they condition on non-missing data from all three files. In this manner, their imputation is based on the complete likelihood function for the linked longitudinal data.

The Abowd and Woodcock technique necessitates some data reduction. To understand the data reduction, consider imputing missing data in the individual characteristics file U. Since individuals have work histories with different dynamic configurations of employers, explicitly conditioning the missing data imputation of individual characteristics on every variable corresponding to each job held by each worker is impractical – there are a different number of such variables for each observation to be imputed. A sensible alternative is to condition on some function of the available data that is well defined for each observation. For example, to impute missing data in U, one could condition on the person-specific means of time-varying work history and firm variables. Similar data reductions are required to impute missing data in the other files. In what follows, we use the functions g, h, m and n to represent data reductions that span sampling frames.

Abowd and Woodcock note the importance of conditioning the imputation of time-varying variables on contemporaneous data and leads and lags of available data. Because the dynamic configuration of work histories varies from worker to worker and the pattern of firm "births" and "deaths" varies from firm to firm, not every observation with missing data has the same number of leads and lags available to condition the imputation. In some cases, there are no leads and lags available at all. They suggest grouping observations by the availability of dynamic conditioning data (i.e., the number of leads and lags available to condition missing data imputations) and separately imputing missing data for each group. This maximizes the set of conditioning variables used to impute each missing value. Again, some data reduction is generally necessary to keep the number of groups reasonable. For example, one might only condition on a maximum of s leads and lags, with $s = 1$ or $s = 2$. They parameterize the set of dynamic conditioning data available for a particular observation by κ_{it} in the work history file, and γ_{jt} in the firm file. It may also be desirable to split the observations into separate groups on the basis of observable characteristics, for example sex, full-time/part-time employment status, or industry. They parameterize these groups by λ_i in the individual file, μ_{it} in the work history file, and ν_{jt} in the firm file.

The key aspects of the SRMI algorithm are as follows. One proceeds sequentially and iteratively through variables with missing data from all three files, at each stage

imputing missing data conditional on all non-missing data and the most recently imputed values of missing data. The optimal imputation sequence is in increasing degree of missingness. As each variable in the sequence comes up for imputation, observations are split into groups based on the value of κ_{it}, γ_{jt}, λ_i, μ_{it}, and/or ν_{jt}. The imputed values are sampled from the posterior predictive distribution of a parametric Bayesian imputation model that is specific to each group. After the imputes are drawn, the source file for the variable under imputation is reassembled from each of the group files. Before proceeding to the next variable, all three files must be updated with the most recent imputations, since the next variable to be imputed may reside in another file (U, W, or Z). At the same time, the functions of conditioning data (including leads and lags) described above generally need to be re-computed. The procedure continues for a pre-specified number of rounds or until the imputed values are stable.

Explicitly specifying the posterior predictive densities from which the imputations are drawn is notationally cumbersome. For completeness, we reproduce these directly from Abowd and Woodcock in (22.1), (22.2), and (22.3). For a particular variable under imputation, subscripted by k, they denote by $U_{<k}$ the set of variables in U with less missing data than variable k; $W_{<k}$ and $Z_{<k}$ are defined analogously. They denote by $U_{>k}$ the set of variables in U with more missing data than variable k, and define $W_{>k}$ and $Z_{>k}$ similarly. They use the subscript *obs* to denote variables with no missing data. They also subscript conditioning variables by i, j, and t as appropriate to make clear the relationships between variables in the three data files. The predictive densities from which the round $\ell+1$ imputations are drawn are

$$\int f_{u_k}\left(u_k \left| \begin{array}{c} U_{<k,i}^{(\ell+1)}, U_{>k,i}^{(\ell)}, U_{\text{obs},i}\ , \\ g_k\left(\left\{Z_{<k,J(i,t)}^{(\ell+1)}, Z_{>k,J(i,t)}^{(\ell)}, Z_{\text{obs},J(i,t)}\right\}_{t=1}^{t=T_i}\right)\ , \\ h_k\left(\left\{W_{<k,it}^{(\ell+1)}, W_{>k,it}^{(\ell)}, W_{\text{obs},it}\right\}_{t=1}^{t=T_i}, \lambda_i, \theta_k\right) \end{array}\right.\right) p_k(\theta_k|.)\,\mathrm{d}\theta_k \tag{22.1}$$

$$\int f_{w_k}\left(w_k \left| \begin{array}{c} U_{<k,i}^{(\ell+1)}, U_{>k,i}^{(\ell)}, U_{\text{obs},i}\ , \\ \left\{Z_{<k,J(i,\tau)}^{(\ell+1)}, Z_{>k,J(i,\tau)}^{(\ell)}, Z_{\text{obs},J(i,\tau)}\right\}_{\tau=t-s}^{\tau=t+s}\ , \\ \left\{w_{k,i\tau}^{(\ell)}\right\}_{\tau=t-s,\tau\neq t}^{\tau=t+s}\ , \\ \left\{W_{<k,i\tau}^{(\ell+1)}, W_{>k,i\tau}^{(\ell)}, W_{\text{obs},i\tau}\right\}_{\tau=t-s}^{\tau=t+s}, \kappa_{it}, \mu_{it}, \theta_k \end{array}\right.\right) p_k(\theta_k|.)\,\mathrm{d}\theta_k \tag{22.2}$$

$$\int f_{z_k}\left(z_k \left| \begin{array}{c} m_k\left(U_{<k,J^{-1}(i,t)}^{(\ell+1)}, U_{>k,J^{-1}(i,t)}^{(\ell)}, U_{\text{obs},J^{-1}(i,t)}\right), \\ \left\{z_{k,j\tau}^{(\ell)}\right\}_{\tau=t-s,\tau\neq t}^{\tau=t+s}, \left\{Z_{<k,j\tau}^{(\ell+1)}, Z_{>k,j\tau}^{(\ell)}, Z_{\text{obs},j\tau}\right\}_{\tau=t-s}^{\tau=t+s}, \\ n_k\left(\left\{\begin{array}{c} W_{<k,J^{-1}(i,\tau)\tau}^{(\ell+1)}, W_{>k,J^{-1}(i,\tau)\tau}^{(\ell)}, \\ W_{\text{obs},J^{-1}(i,\tau)\tau} \end{array}\right\}_{\tau=t-s}^{\tau=t+s}, \right. \\ \left. \gamma_{jt}, \nu_{jt}, \theta_k\right) \end{array}\right.\right) p_k(\theta_k|.)\,\mathrm{d}\theta_k, \tag{22.3}$$

where $f_{.k}$ is the likelihood defined by an appropriate generalized linear model for variable k, θ_k are unknown parameters, and the posterior densities $p_k(\theta_k|.)$ are conditioned on the same information as $f_{.k}$. Repeating the missing data imputation method M times yields M sets of completed data files (U^m, W^m, Z^m) which they call the completed data implicates $m = 1, \ldots, M$.

Equations (22.1–22.3) describe the complete set of conditional distributions of each variable in the linked longitudinal employer–employee data, given all other variables. Hence, they form the basis for sampling from this complete distribution. One can use these equations in a Gibbs sampler or other Monte Carlo Markov Chain algorithm to draw a complete sample of linked longitudinal data that has the same likelihood function as the original analysis sample. Abowd and Woodcock use this property to draw partially synthetic data from the joint posterior predictive distribution.

22.2.2 Sampling from Linked Data

Many of the estimators discussed below are computationally intensive. Because many longitudinal linked databases are constructed from administrative records they are very large.[4] Thus researchers are sometimes faced with the prospect of sampling from the linked data to facilitate estimation. In principle, sampling from any one of the frames (workers, firms, or jobs) that comprise the linked data is straightforward. However, the estimators discussed below rely on links between sampling frames (i.e., observed worker mobility between firms) for identification. Small simple random samples of individuals may not retain sufficient "connectedness" between sampling frames for identification.[5]

Woodcock (2003) considers the problem of sampling from linked data while preserving a minimum degree of connectedness between sampling frames. He presents a "dense" sampling algorithm that guarantees each sampled worker is connected to at least n others by a common employer. The sample is otherwise representative of the population of individuals employed in a reference period. The dense sampling algorithm is straightforward. It operates on the population of jobs at firms with at least n employees in the reference period t. In the first stage, sample firms with probabilities proportional to their employment in period t. In the second stage, sample a minimum of n employees from each sampled firm, with probabilities inversely proportional to the firm's employment in period t. A simple application of Bayes' rule demonstrates that all jobs active in period t have an equal probability of being sampled. The sample is thus equivalent to a simple random sample of jobs active in period t, but guarantees that each sampled worker is connected to at least n others.

[4] See Abowd and Kramarz (1999a) for a typology.

[5] See Sect. 22.5.1.2 below for a discussion of connectedness and its role in identifying person and firm effects.

22.3 Linear Statistical Models with Person and Firm Effects

22.3.1 A General Specification

We consider the general linear statistical model:

$$y_{it} = x_{it}\beta + q_{it,\mathrm{J}(i,t)}\theta_i + r_{it,\mathrm{J}(i,t)}\psi_{\mathrm{J}(i,t)} + \varepsilon_{it} \tag{22.4}$$

where y_{it} is an observation for individual $i = 1,\ldots,N$, $t = n_{i1},\ldots,n_{iT_i}$, T_i is the total number of periods of data available for individual i, and the indices $n_{i1},\ldots,n_{iT_i}$ indicate the period corresponding to the first observation on individual i through the last observation on that individual, respectively. The vectors x_{it} contain P time-varying, exogenous characteristics of individual i; the vectors $q_{it,\mathrm{J}(i,t)}$, and $r_{it,\mathrm{J}(i,t)}$ contain respectively Q and R exogenous characteristics of individual i and (or) firm $\mathrm{J}(i,t)$. Both vectors include indicators that associate an observation and a person (for q) or a firm (for r). We denote the design matrices of these indicators by D and F, respectively. The vector θ_i is a size Q vector of person effects; $\psi_{\mathrm{J}(i,t)}$ is a size R vector of firm effects; and ε_{it} is the statistical residual. The first period available for any individual is arbitrarily dated 1 and the maximum number of periods of data available for any individual is T. Assemble the data for each person i into conformable vectors and matrices

$$y_i = \begin{bmatrix} y_{i,n_{i1}} \\ \cdots \\ y_{i,n_{iT_i}} \end{bmatrix},$$

$$X_i = \begin{bmatrix} x_{i,n_{i1},1} & \cdots & x_{i,n_{i1},P} \\ & \cdots & \\ x_{i,n_{iT_i},1} & \cdots & x_{i,n_{iT_i},P} \end{bmatrix},$$

$$\varepsilon_i = \begin{bmatrix} \varepsilon_{i,n_{i1}} \\ \cdots \\ \varepsilon_{i,n_{iT_i}} \end{bmatrix}$$

where y_i and ε_i are $T_i \times 1$ and X_i is $T_i \times P$ with similar definitions for $Q_{i,\mathrm{J}(i,.)}$ and $R_{i,\mathrm{J}(i,.)}$.

We assume that a simple random sample of N individuals is observed for a maximum of T periods. Assume further that ε_i has the following properties:

$$\mathrm{E}\left[\varepsilon_i | X_i, Q_{i,\mathrm{J}(i,.)}, R_{i,\mathrm{J}(i,.)}\right] = 0$$

and

$$\mathrm{Cov}\left[\varepsilon_i, \varepsilon_m | X_i, Q_{i,\mathrm{J}(i,.)}, R_{i,\mathrm{J}(i,.)}, X_m, Q_{m,\mathrm{J}(m,.)}, R_{m,\mathrm{J}(m,.)}\right]$$
$$= \begin{cases} \{\Sigma_{T_i}\}_i, i = m \\ 0, \text{otherwise} \end{cases}$$

where $\{\Sigma_{T_i}\}_i$ means the selection of rows and columns from a $T \times T$ positive definite symmetric matrix Σ such that the resulting $T_i \times T_i$ positive definite symmetric matrix corresponds to the periods $\{n_{i1}, n_{i2}, \ldots, n_{iT_i}\}$.[6] In full matrix notation we have

$$y = X\beta + \left[D, \widetilde{Q}\right]\theta + \left[F, \widetilde{R}\right]\psi + \varepsilon \tag{22.5}$$

where: X is the $N^* \times P$ matrix of observable, time-varying characteristics (in deviations from the grand means); D is the $N^* \times N$ design matrix of indicator variables for the individual; $\widetilde{Q}$ is the $N^* \times (Q-1)N$ matrix of the observable characteristics in q with person-specific effects; F is the $N^* \times J$ design matrix of indicator variables for the firm; $\widetilde{R}$ is the $N^* \times (R-1)J$ matrix of observable characteristics in r with firm-specific effects; y is the $N^* \times 1$ vector of dependent data (also in deviations from the grand mean); ε is the conformable vector of residuals; and $N^* = \sum_{i=1}^{N} T_i$. The vector y is ordered according to individuals as

$$y = \begin{bmatrix} y_1 \\ \cdots \\ y_N \end{bmatrix} \tag{22.6}$$

and X, Q, R and ε are ordered conformably. A typical element of y is y_{it} and a typical element of X, or any similarly organized matrix, as $x_{(i,t)p}$ where the pair (i,t) denotes the row index and p denotes the column index. The effects in (22.4) and (22.5) are: β, the $P \times 1$ vector of coefficients on the time-varying personal characteristics; θ, the $QN \times 1$ vector of individual effects; and ψ, the $RJ \times 1$ vector of firm effects. When estimating the model by fixed effects methods, identification of the effects is accomplished by imposing a zero sample mean for θ_i and $\psi_{\mathrm{J}(i,t)}$ taken over all (i,t).[7] In the mixed effects case, identification is achieved by assuming the random effects have zero conditional mean and finite conditional variance.

22.3.2 The Pure Person and Firm Effects Specification

A simpler specification is:

$$y_{it} = x_{it}\beta + \theta_i + \psi_{\mathrm{J}(i,t)} + \varepsilon_{it} \tag{22.7}$$

with variables defined as above except that θ_i is the *pure person effect* and $\psi_{\mathrm{J}(i,t)}$ is the *pure firm effect*. We now assume that ε_i has the following properties:[8]

[6] See Sect. 22.6 for a specific example of $\{\Sigma_{T_i}\}_i$.

[7] Further details of identification requirements are discussed in Sect. 22.5.1.

[8] The zero conditional mean assumption (22.8) has been interpreted as an assumption of "exogenous mobility," since it precludes any relationship between an individuals employment location (measured by F_i) and the errors ε_i. See AKM for further discussion, and Sect. 22.8 below for recent work that accomodates endogenous mobility.

$$\mathrm{E}[\varepsilon_i | D_i, F_i, X_i] = 0 \tag{22.8}$$

and

$$\mathrm{Cov}[\varepsilon_i, \varepsilon_m | D_i, D_m, F_i, F_m, X_i, X_m] = \begin{cases} \{\Sigma_{T_i}\}_i, & i = m \\ 0, & \text{otherwise} \end{cases}$$

where D_i and F_i are those elements of D and F, respectively, corresponding to person i. In full matrix notation we have

$$y = X\beta + D\theta + F\psi + \varepsilon \tag{22.9}$$

where: X is the $N^* \times P$ matrix of observable, time-varying characteristics (in deviations from the grand means); D is the $N^* \times N$ design matrix of indicator variables for the individual; F is the $N^* \times J$ design matrix of indicator variables for the employer at which i works at date t (J firms total); y is the $N^* \times 1$ vector of dependent data (also in deviations from the grand mean); ε is the conformable vector of residuals; and $N^* = \sum_{i=1}^{N} T_i$.

The effects in (22.7) and (22.9) are: β, the $P \times 1$ vector of coefficients on the time-varying personal characteristics; θ, the $N \times 1$ vector of individual effects; and ψ, the $J \times 1$ vector of firm effects. As above, identification of the effects is accomplished by imposing a zero sample mean for θ_i and $\psi_{\mathrm{J}(i,t)}$ taken over all (i,t) for fixed-effects estimators, and by assuming of zero conditional mean and finite conditional variance for random-effects estimators.

22.4 Definition of Effects of Interest

Many familiar models are special cases of the linear model in (22.4) and (22.5) or the simpler version in (22.7) and (22.9). In this section we define a variety of effects of interest that are functions of the person and firm effects specified in the preceding section. These definitions allow us to consider these familiar models using common notation and internally coherent definitions. We use the example of estimating inter-industry wage differentials, frequently called industry effects, to illustrate some important issues.

22.4.1 Person Effects and Unobservable Personal Heterogeneity

The person effect in (22.7) combines the effects of observable time-invariant personal characteristics and unobserved personal heterogeneity. We decompose these two parts of the pure person effect as

$$\theta_i = \alpha_i + u_i\eta \tag{22.10}$$

where α_i is the unobservable personal heterogeneity, u_i is a vector of time-invariant personal characteristics, and η is a vector of effects associated with the time-invariant personal characteristics. An important feature of the decomposition in (22.10) is that estimation can proceed for the person effects, θ_i, whether random or fixed, without direct estimation of η. Since many linked employer–employee data sets contain limited, or missing, information on the time-invariant characteristics u_i, we describe the estimation algorithms in terms of θ_i; however, when data on u_i are available, equivalent techniques can be used for estimation in the presence of α_i (see AKM for the fixed effects case, Woodcock (2003) for the mixed effects case). The design matrix D in (22.9) can be augmented by columns associated with the observables u_i so that the statistical methods discussed below are applicable to the estimation of the effect specified in (22.10).

This specification can be further generalized by incorporating time-varying observable characteristics of the worker, q_{it}, or of the firm, q_{jt}, that may well be interacted as in (22.4) and (22.5) to give:

$$\theta_{jit} = \alpha_i + u_i\eta + q_{it}\mu_i + q_{jt}\delta_i \tag{22.11}$$

where μ_i and δ_i are vectors of effects associated with the time-varying person and firm observable characteristics. Statistical analysis of the effects defined by (22.11) is accomplished by augmenting the columns of D to reflect the data in q_{jt} and q_{it}. The formulae shown in the estimation sections below can then be applied to the augmented design matrix.

22.4.2 Firm Effects and Unobservable Firm Heterogeneity

The firm effect in (22.7) combines the effects of observable and unobserved time-invariant characteristics of the firm. It can also be generalized to contain the effects of time-varying characteristics of the firm and time-varying characteristics of the employee–employer match as in (22.4) and (22.5). We illustrate each of these possibilities in this subsection.

We can decompose the pure firm effect of (22.7) into observable and unobservable components as

$$\psi_j = \phi_j + v_j\rho \tag{22.12}$$

where ϕ_j is unobservable firm heterogeneity, v_j is a vector of time-invariant firm characteristics, and ρ is a vector of associated effects.

Time-varying firm and employer–employee match characteristics require a redefinition of the simple firm effect as ψ_{jit}. The addition of the i and t subscripts allows the firm effect to vary over time and across employer–employee matches. Now let the firm observable characteristics be time-varying, v_{jt}, and denote the observable match characteristics by r_{jit}. Then we can write the firm effect as

$$\psi_{jit} = \phi_j + v_{jt}\rho + r_{jit}\gamma_j \tag{22.13}$$

where γ_j is a vector of effects associated with the match characteristics. Statistical analysis of the effects defined by equation (22.13) is accomplished by augmenting the columns of F to reflect the data in v_{jt} and r_{jit}. The formulas shown in the estimation sections below can then be applied to the augmented design matrix.

22.4.3 Firm-Average Person Effect

For each firm j we define a firm-average person effect

$$\bar{\theta}_j \equiv \bar{\alpha}_j + \bar{u}_j \eta = \frac{\sum_{\{(i,t)|\mathrm{J}(i,t)=j\}} \theta_i}{N_j} \tag{22.14}$$

where

$$N_j \equiv \sum_{\forall(i,t)} 1(\mathrm{J}(i,t) = j)$$

and the function $1(A)$ takes the value 1 if A is true and 0 otherwise. The importance of the effect defined in (22.14) may not be apparent at first glance. Consider the difference between ψ_j and $\bar{\theta}_j$. The former effect measures the extent to which firm j deviates from the average firm (averaged over individuals and weighted by employment duration) whereas the latter effect measures the extent to which the average employee of firm j deviates from the population of potential employees. In their analysis of wage rate determination, AKM refer to the firm-average person effect, $\bar{\theta}_j$, as capturing the idea of high (or low) wage workers while the pure firm effect, ψ_j, captures the idea of a high (or low) wage firm. Both effects must be specified and estimable for the distinction to carry empirical import.

22.4.4 Person-Average Firm Effect

For each individual i consider the person-average firm effect defined as

$$\bar{\psi}_i \equiv \bar{\phi}_i + \bar{v}_i \rho = \frac{\sum_t \psi_{\mathrm{J}(i,t)it}}{T_i}. \tag{22.15}$$

This effect is the individual counterpart to the firm-average person effect. Limited sample sizes for individuals make estimates of this effect less useful in their own right; however, they form the basis for conceptualizing the difference between the effect of heterogeneous individuals on the composition of a firm's workforce, as measured by the effect defined in (22.14), and the effect of heterogeneous firms on an individual's career employment outcomes, as measured by the effect in (22.15).

22.4.5 Industry Effects[9]

Industry is a characteristic of the employer. As such, the analysis of industry effects in the presence of person and firm effects can be accomplished by appropriate definition of the industry effect with respect to the firm effects. We call the properly defined industry effect a "pure" industry effect. Denote the pure industry effect, conditional on the same information as in (22.7) and (22.9), as κ_k for some industry classification $k = 1, \ldots, K$. Our definition of the pure industry effect is simply the correct aggregation of the pure firm effects within the industry. We define the pure industry effect as the one that corresponds to putting industry indicator variables in equation (22.9) and, then, defining what is left of the pure firm effect as a deviation from the industry effects. Hence, κ_k can be represented as an employment-duration weighted average of the firm effects within the industry classification k:

$$\kappa_k \equiv \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\frac{1(\mathrm{K}(\mathrm{J}(i,t)) = k)\psi_{\mathrm{J}(i,t)}}{N_k} \right]$$

where

$$N_k \equiv \sum_{j=1}^{J} 1(\mathrm{K}(j) = k) N_j$$

and the function $\mathrm{K}(j)$ denotes the industry classification of firm j. If we insert this pure industry effect, the appropriate aggregate of the firm effects, into (22.7), then

$$y_{it} = x_{it}\beta + \theta_i + \kappa_{\mathrm{K}(\mathrm{J}(i,t))} + (\psi_{\mathrm{J}(i,t)} - \kappa_{\mathrm{K}(\mathrm{J}(i,t))}) + \varepsilon_{it}$$

or, in matrix notation as in (22.9),

$$y = X\beta + D\theta + FA\kappa + (F\psi - FA\kappa) + \varepsilon \tag{22.16}$$

where the matrix A, $J \times K$, classifies each of the J firms into one of the K industries; that is, $a_{jk} = 1$ if, and only if, $\mathrm{K}(j) = k$. Algebraic manipulation of (22.16) reveals that the vector κ, $K \times 1$, may be interpreted as the following weighted average of the pure firm effects:

$$\kappa \equiv (A'F'FA)^{-1}A'F'F\psi. \tag{22.17}$$

and the effect $(F\psi - FA\kappa)$ may be re-expressed as $M_{FA}F\psi$, where $M_Z \equiv I - Z(Z'Z)^{-}Z'$ denotes the column null space of an arbitrary matrix Z, and $()^{-}$ is a computable generalized inverse. Thus, the aggregation of J firm effects into K industry effects, weighted so as to be representative of individuals, can be accomplished directly by the specification of (22.16). Only $\mathrm{rank}(F'M_{FA}F)$ firm effects can be separately identified using unrestricted fixed-effects methods; however, there is neither an omitted variable nor an aggregation bias in the estimates of (22.16), using either of class of estimators discussed below. Equation (22.16) simply decomposes $F\psi$ into two orthogonal components: the industry effects $FA\kappa$, and what is left of the

[9] This section is based upon the analysis in Abowd, Finer and Kramarz (1999).

firm effects after removing the industry effect, $M_{FA}F\psi$. While the decomposition is orthogonal, the presence of X and D in (22.16) greatly complicates the estimation by either fixed-effects or mixed-effects techniques.

22.4.6 Other Firm Characteristic Effects

Through careful specification of the firm effect in (22.13), we can estimate the average effect associated with any firm characteristic, v_{jt}, or any interaction of firm and personal characteristics, r_{jit}, while allowing for unobservable firm and personal heterogeneity.

22.4.7 Occupation Effects and Other Person × Firm Interactions

If occupation effects are interpreted as characteristics of the person, then they are covered by the analysis above and can be computed as functions of θ as described in (22.11). Occupation effects are often interpreted as an interaction between person and firm effects (Groshen (1991a,b, implicitly). Mixed effects specifications are most appropriate in this case, and are discussed in Sect. 22.6.

22.5 Estimation by Fixed Effects Methods

In this section we present methods for estimating the pure person and firm effects specification (22.7) by direct least squares, and consistent methods for estimating generalizations of this specification.

22.5.1 Estimation of the Fixed Effects Model by Direct Least Squares

This subsection directly draws from Abowd, Creecy and Kramarz (2002) (ACK, hereafter). The normal equations for least squares estimation of fixed person, firm, and characteristic effects are of very high dimension. Thus estimating the full model by fixed-effect methods requires special algorithms. In our earlier work, e.g., Abowd, Finer and Kramarz (1999) (AFK, hereafter) and AKM, we relied on statistical approximations to render the estimation problem tractable. More recently, ACK developed new algorithms that permit the exact least squares estimation of all the effects in (22.7). These algorithms are based on the iterative conjugate gradient method and rely on computational simplifications admitted by the sparse structure

of the least squares normal equations. They have some similarity to methods used in the animal and plant breeding literature.[10] ACK also developed new methods for computing estimable functions of the parameters of (22.7).

22.5.1.1 Least Squares Normal Equations

The full least squares solution to the estimation problem for (22.7) solves the normal equations for all estimable effects:

$$\begin{bmatrix} X'X & X'D & X'F \\ D'X & D'D & D'F \\ F'X & F'D & F'F \end{bmatrix} \begin{bmatrix} \beta \\ \theta \\ \psi \end{bmatrix} = \begin{bmatrix} X'y \\ D'y \\ F'y \end{bmatrix} \tag{22.18}$$

In typical applications, the cross-product matrix on the left-hand side of the equation is too high-dimensional to solve using conventional algorithms (e.g., those implemented in SAS, Stata, and other general purpose linear modeling software based on variations of the sweep algorithm for solving (22.18)). AKM present a set of approximate solutions based on the use of different conditioning effects, Z. AFK applies the best of these approximations with a much higher-dimension Z.

22.5.1.2 Identification of Individual and Firm Effects

Many interesting economic applications of (22.7) make use of the estimated person and firm effects. Estimation requires a method for determining the identified effects[11]. The usual technique of sweeping out singular row/column combinations from the normal (22.18) is not applicable to the ACK method because they solve the normal equations without inverting the cross-product matrix. Hence, identification requires finding conditions under which the normal equations (22.18) can be solved exactly for some estimable functions of the person and firm effects. In this subsection we ignore the problem of identifying the coefficients β because in practice this is rarely difficult.

The identification problem for the person and firm effects can be solved by applying graph-theoretic methods to determine groups of connected individuals and firms. Within a connected group of persons/firms, identification can be determined using conventional methods from the analysis of covariance. Connecting persons and firms requires that some of the individuals in the sample be employed at multiple

[10] See Abowd and Kramarz (1999a) for a longer dicussion of the relation of these models to those found in the breeding literature. The techniques are summarized in Robinson (1991) and the random-effects methods are thoroughly discussed in Neumaier and Groeneveld (1996). The programs developed for breeding applications cannot be used directly for the linked employer–employee data application because of the way the breeding effects are parameterized.

[11] Standard statistical references, for example Searle et al. (1992), provide general methods for finding the estimable functions of the parameters of (22.7). These methods also require the solution of a very high dimension linear system and are, therefore, impractical for our purposes.

employers. When a group of persons and firms is connected, the group contains all the workers who ever worked for any of the firms in the group and all the firms at which any of the workers were ever employed. In contrast, when a group of persons and firms is not connected to a second group, no firm in the first group has ever employed a person in the second group, nor has any person in the first group ever been employed by a firm in the second group. From an economic perspective, connected groups of workers and firms show the realized mobility network in the economy. From a statistical perspective, connected groups of workers and firms block-diagonalize the normal equations and permit the precise statement of identification restrictions on the person and firm effects.

The following algorithm constructs G mutually-exclusive groups of connected observations from the N workers in J firms observed over the sample period.[12]

```
For g = 1,..., repeat until no firms remain:
  The first firm not assigned to a group is in group g.
  Repeat until no more firms or persons are added to
    group g:
    Add all persons employed by a firm in group g
    to group g.
    Add all firms that have employed a person in group g
    to group g.
  End repeat.
End for.
```

At the conclusion of the algorithm, the persons and firms in the sample have been divided into G groups. Denote the number of individuals in group g by N_g, and the number of employers in the group by J_g. Some groups contain a single employer and, possibly, only one individual. For groups that contain more than one employer, every employer in the group is connected (in the graph-theoretic sense) to at least one other employer in the group. Within each group g, the group mean of y and $N_g - 1 + J_g - 1$ person and firm effects are identified. After the construction of the G groups, exactly $N + J - G$ effects are estimable. See the proof in Appendix 1 of ACK.[13]

22.5.1.3 Normal Equations after Group Blocking

The identification argument can be clarified by considering the normal equations after reordering the persons and firms by group. For simplicity, let the arbitrary

[12] This algorithm finds all of the maximally connected sub-graphs of a graph. The relevant graph has a set of vertices that is the union of the set of persons and the set of firms and edges that are pairs of persons and firms. An edge (i, j) is in the graph if person i has worked for firm j.

[13] The grouping algorithm constructs groups within which "main effect" contrasts due to persons and firms are identified. In the linear models literature the "groups" are called "connected data". See Searle (1987, pp. 139–149) for a discussion of connected data. See Weeks and Williams (1964) for the general algorithm in analysis of variance models.

equation determining the unidentified effect set it equal to zero, i.e., set one person or firm effect equal to zero in each group. Then the column associated with this effect can be removed from the reorganized design matrix and we can suppress the column associated with the group mean. The resulting normal equations are:

$$\left[\begin{array}{c|cc|cc|c|cc} X'X & X'D_1 & X'F_1 & X'D_2 & X'F_2 & \cdots & X'D_G & X'F_G \\ \hline D_1'X & D_1' & D_1 & 0 & 0 & \cdots & 0 & 0 \\ F_1'X & F_1' & D_1 & 0 & 0 & \cdots & 0 & 0 \\ \hline D_2'X & 0 & 0 & D_2'D_2 & D_2'F_2 & \cdots & 0 & 0 \\ F_2'X & 0 & 0 & F_2'D_2 & F_2'F_2 & \cdots & 0 & 0 \\ \hline \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \hline D_G'X & 0 & 0 & 0 & 0 & \cdots & D_G'D_G & D_G'F_G \\ F_G'X & 0 & 0 & 0 & 0 & \cdots & F_G'D_F & F_G'F_G \end{array}\right] \left[\begin{array}{c} \beta \\ \hline \theta_1 \\ \psi_1 \\ \hline \theta_2 \\ \psi_2 \\ \hline \cdots \\ \hline \theta_G \\ \psi_G \end{array}\right] = \left[\begin{array}{c} X'y \\ \hline D_1'y \\ F_1'y \\ \hline D_2'y \\ F_2'y \\ \hline \cdots \\ \hline D_G'y \\ F_G'y \end{array}\right] \tag{22.19}$$

After reordering by group, the cross-products matrix is block diagonal. This matrix has full column rank and the solution for the parameter vector is unique. ACK do not solve (22.19) directly. Rather, they apply the technique discussed below to estimate the identifiable effects.

22.5.1.4 Estimation by Direct Solution of the Least Squares Problem

Appendix 2 in ACK shows the exact algorithm used to solve equation (22.18). It is a variant of the conjugate gradient algorithm, customized to exploit the sparse representation of (22.18) and to accommodate very large problems with many X variables. In practice, ACK apply this algorithm to the full set of persons, firms and characteristics shown in the design matrices of (22.7) and (22.18). Unlike (22.19), the cross-product matrix in (22.18) is not of full rank. Although the algorithm ACK use converges to a least squares solution, the parameter estimates are not unique. They subsequently apply the following identification procedure to the estimated effects. In each group, they eliminate one person effect by normalizing the group mean person effect to zero. ACK also normalize the overall mean person and firm effects to zero. This procedure identifies the grand mean of the dependent variable (or the overall regression constant if X and y have not been standardized to mean zero) and a set of $N+J-G-1$ person and firm effects measured as deviations from the grand mean of the dependent variable.[14]

[14] The computer software is available from the authors for both the direct least squares estimation of the two-factor analysis of covariance and the grouping algorithm. Computer software that implements both the random and fixed effects versions of these models used in breeding applications can be found in Groeneveld (1998). The specific algorithm we use can be found in Dongarra et al. (1991) p. 146.

22.5.2 Consistent Methods for β and γ (The Firm-Specific Returns to Seniority)

The preceding discussion focused on estimation of the pure person and firm effects model (22.7). In this subsection, we discuss methods presented in AKM for consistent estimation of more general representations of the person and firm effects. In particular, we discuss consistent estimation of β and γ_j in the general representation of the firm effect (22.13). The method relies on within-individual-firm differences of the data. It is robust in the sense that it requires no additional statistical assumptions beyond those specified in (22.4) and the general definition of the firm effect (22.13).[15] We should note, however, that this estimation technique relies heavily on the assumption of no interaction between X and F. Consider the first differences:

$$y_{i,n_{it}} - y_{in_{it-1}} = (x_{in_{it}} - x_{in_{it-1}})\beta + \gamma_{\mathrm{J}(i,n_{it})}(s_{in_{it}} - s_{in_{it-1}}) + \varepsilon_{in_{it}} - \varepsilon_{in_{it-1}} \tag{22.20}$$

for all observations for which $\mathrm{J}(i,n_{it}) = \mathrm{J}(i,n_{it-1})$, and where $s_{in_{it}}$ represents worker i's seniority at firm $J(i,n_{it})$ in period n_{it}.[16] In matrix form:

$$\triangle y = \triangle X\beta + \widetilde{F}\gamma + \triangle\varepsilon \tag{22.21}$$

where $\triangle y$ is $\widetilde{N^*} \times 1$, $\triangle X$ is $\widetilde{N^*} \times P$, $\widetilde{F}$ is $\widetilde{N^*} \times J$, $\triangle\varepsilon$ is $\widetilde{N^*} \times 1$, and $\widetilde{N^*}$ is equal to the number of (i,t) combinations in the sample that satisfy the condition $\mathrm{J}(i,n_{it}) = \mathrm{J}(i,n_{it-1})$. The matrix $\widetilde{F}$ is the rows of the design of γ that correspond to the person-years (i,t) for which the condition $\mathrm{J}(i,n_{it}) = \mathrm{J}(i,n_{it-1})$ is satisfied. The least squares estimates of β and γ are,

$$\widetilde{\beta} = (\triangle X' M_{\widetilde{F}} \triangle X)^{-1} \triangle X' M_{\widetilde{F}} \triangle y \tag{22.22}$$

$$\widetilde{\gamma} = (\widetilde{F}'\widetilde{F})^{-1}\widetilde{F}'(\triangle y - \triangle X\widetilde{\beta}). \tag{22.23}$$

A consistent estimate of $\mathrm{V}[\widetilde{\beta}]$ is given by

$$\widetilde{\mathrm{V}[\widetilde{\beta}]} = (\triangle X' M_{\widetilde{F}} \triangle X)^{-1}(\triangle X' M_{\widetilde{F}} \widetilde{\Omega} M_{\widetilde{F}} \triangle X)(\triangle X' M_{\widetilde{F}} \triangle X)^{-1}$$

where

$$\widetilde{\Omega} \equiv \begin{bmatrix} \widetilde{\Omega}[\triangle\varepsilon_1] & 0 & \cdots & 0 \\ 0 & \widetilde{\Omega}[\triangle\varepsilon_2] & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \widetilde{\Omega}[\triangle\varepsilon_{N^*}] \end{bmatrix}$$

[15] We have excluded $v_{jt}\rho$ from the firm effect (22.13), and assume a pure person effect θ_i.

[16] In our preceding notation for the general firm effect (22.13), seniority is an element of observable match-specific characteristics r_{ijt}.

and

$$\widetilde{\Omega}[\triangle\varepsilon_i] \equiv \begin{bmatrix} \widetilde{\triangle\varepsilon}_{in_2}^2 & \widetilde{\triangle\varepsilon}_{in_2}\widetilde{\triangle\varepsilon}_{in_3} & \cdots & \widetilde{\triangle\varepsilon}_{in_2}\widetilde{\triangle\varepsilon}_{in_{T_i}} \\ \widetilde{\triangle\varepsilon}_{in_3}\widetilde{\triangle\varepsilon}_{in_2} & \widetilde{\triangle\varepsilon}_{in_3}^2 & \cdots & \widetilde{\triangle\varepsilon}_{in_3}\widetilde{\triangle\varepsilon}_{in_{T_i}} \\ \cdots & \cdots & \cdots & \cdots \\ \widetilde{\triangle\varepsilon}_{in_{T_1}}\widetilde{\triangle\varepsilon}_{in_2} & \widetilde{\triangle\varepsilon}_{in_{T_1}}\widetilde{\triangle\varepsilon}_{in_3} & \cdots & \widetilde{\triangle\varepsilon}_{in_{T_i}}\widetilde{\triangle\varepsilon}_{in_{T_i}} \end{bmatrix}.$$

It is understood that only the rows of $\triangle\varepsilon$ that satisfy the condition $\mathrm{J}(i,n_{it}) = \mathrm{J}(i,n_{it-1})$ are used in the calculation of $\widetilde{\Omega}$, which is therefore $\widetilde{N^*} \times \widetilde{N^*}$. Notice that this estimator does not impose all of the statistical structure of the basic linear model (22.7).

22.6 The Mixed Model

In this Section, we focus on a mixed model specification of the pure person and firm effects model. The mixed model arises when some, or all, of the effects in (22.9) are treated as random, rather than fixed, effects. There is considerable confusion in the literature about the comparison of fixed and mixed effects specifications, and so we take pains in this section to define terms in a manner consistent with the enormous statistical literature on this subject.

Consider the matrix formulation of the pure person and firm effects model, given in (22.9). We focus on the cases treated by Woodcock (2003) and Abowd and Stinson (2003), where the parameters β on observable characteristics are treated as fixed, and where the pure person and firm effects θ and ψ are random.[17] This specification corresponds closely to the hierarchical models that are common in some other applied settings, for instance in the education literature.[18]

The mixed model is completely specified by (22.9) and the stochastic assumptions[19]

[17] In fact, Woodcock (2003) decomposes the pure person effect θ_i into observable $(u_i\eta)$ and unobserved components (α_i) as in equation (22.10). He treats η as fixed and α_i as random. For clarity of exposition we focus here on the simpler case where θ_i is random.

[18] In the education literature, schools are analogous to firms and students are analogous to workers. Because education data typically exhibit far less mobility (of students between schools) than we observe in labor market data, the usual specification nests student effects within school effects. The analogous hierarchical specification is therefore $y_{it} = x_{it}\beta + \theta_{ij} + \psi_j + \varepsilon_{it}$, where θ_{ij} is the person effect (nested within firm), and where ψ_j and θ_{ij} are specified as random effects. Dostie (2005) and Lillard (1999) estimate related mixed effects specifications for wages where the firm effect is nested within individuals, e.g., $y_{it} = x_{it}\beta + \theta_i + \psi_{ij} + \varepsilon_{it}$.

[19] In general, statisticians do not explicitly condition these expectations on X because they are primarily concerned with experimental data, where X constitutes part of the experimental design. Econometricians, however, are most often confronted with observational data. In this setting, X can rarely be considered a fixed component of the experimental design.

$$E[\theta|X] = E[\psi|X] = E[\varepsilon|D,F,X] = 0 \tag{22.24}$$

$$\text{Cov}\left[\begin{array}{c|c} \theta & \\ \psi & X \\ \varepsilon & \end{array}\right] = \begin{bmatrix} \sigma_\theta^2 I_N & 0 & 0 \\ 0 & \sigma_\psi^2 I_J & 0 \\ 0 & 0 & R \end{bmatrix}. \tag{22.25}$$

It is worth noting that unlike some random effects specifications encountered elsewhere in the econometric literature, the mixed model we have specified does not assume that the design of the random effects (D and F) is orthogonal to the design (X) of the fixed effects (β). Such an assumption is almost always violated in economic data.

A variety of parameterizations of the residual covariance R are computationally feasible. Woodcock (2003) considers several in detail. Abowd and Stinson (2003) consider two more in the context of specifications that allow for multiple jobs in the same (i,t) pair and multiple measures of the dependent variable. The simplest parameterization is $R = \sigma_\varepsilon^2 I_{N^*}$. This specification is useful for making comparisons with the fixed-effect estimation procedure.

The most general parameterization estimated by Woodcock (2003) allows for a completely unstructured residual covariance within a worker-firm match. Let M denote the number of worker-firm matches (jobs) in the data, and let $\bar{\tau}$ denote the maximum observed duration of a worker-firm match. Suppose the data are ordered by t within j within i. In the balanced data case, where there are $\bar{\tau}$ observations on each worker-firm match, we can write

$$R = I_M \otimes W \tag{22.26}$$

where W is the $\bar{\tau} \times \bar{\tau}$ within-match error covariance.[20] The extension to unbalanced data, where each match between worker i and firm j has duration $\tau_{ij} \leq \bar{\tau}$, is fairly straightforward. Define a $\bar{\tau} \times \tau_{ij}$ selection matrix S_{ij} with elements on the principal diagonal equal to 1, and off-diagonal elements equal to zero.[21] S_{ij} selects those rows and columns of W that correspond to observed earnings outcomes in the match between worker i and firm j. Then in the unbalanced data case, we have

[20] Woodcock (2003) estimates this parameterization of R under the assumption that W is symmetric and positive semi-definite.

[21] For example, if $\bar{\tau} = 3$ and a match between worker i and firm j lasts for 2 periods,

$$S_{ij} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

$$R = \begin{bmatrix} S'_{11}WS_{11} & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \ddots & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & S'_{1J_1}WS_{1J_1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & S'_{N1}WS_{N1} & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & S'_{NJ_N}WS_{NJ_N} \end{bmatrix}. \quad (22.27)$$

22.6.1 REML Estimation of the Mixed Model

Mixed model estimation is discussed at length in Searle et al. (1992) and McCulloch and Searle (2001). There are three principal methods that can be applied to estimate the variance components $\left(\sigma^2_\theta, \sigma^2_\psi\right)$ and R : ANOVA, Maximum Likelihood (ML), and Restricted Maximum Likelihood (REML). ANOVA and ML methods are familiar to most economists; REML less so.[22] Since REML is by far the most commonly used estimation method among statisticians, it is worth giving it a brief treatment.

REML is frequently described as maximizing that part of likelihood that is invariant to the fixed effects (e.g., β). More precisely, REML is maximum likelihood on linear combinations of the dependent variable y, chosen so that the linear combinations do not contain any of the fixed effects. As Searle et al. (1992, pp. 250–251) show, these linear combinations are equivalent to residuals obtained after fitting the fixed portion of the model (e.g., $X\beta$) via least squares.[23] The linear combinations $k'y$ are chosen so that

$$k'X\beta = 0 \ \forall\beta \quad (22.28)$$

which implies

$$k'X = 0. \quad (22.29)$$

Thus k' projects onto the column null space of X, and is therefore

$$k' = c'\left[I_{N^*} - X\left(X'X\right)^- X'\right] \quad (22.30)$$

$$\equiv c'M_X \quad (22.31)$$

for arbitrary c', and where A^- denotes the generalized inverse of A. When X has rank $r \leq p$, there are only $N^* - r$ linearly independent vectors k' satisfying (22.28).

[22] REML estimation of mixed models is commonplace in statistical genetics and in the plant and animal breeding literature. In recent years, REML has in fact become the mixed model estimation method of choice in these fields, superceding ML and ANOVA.

[23] Note this exercise is heuristic and serves only to motivate the REML approach. Under the stochastic assumptions (22.24) and (22.25), the least squares estimator of β is not BLUE. The BLUE of β is obtained by solving the mixed model equations (22.35).

Define $K' = C'M_X$ with rows k' satisfying (22.28), and where K' and C' have full row rank $N^* - r$. REML estimation of the variance parameters is maximum likelihood on $K'y$ under normality. For $y \sim \mathcal{N}(X\beta, \mathbf{V})$ it follows that

$$K'y \sim N\left(0, K'\mathbf{V}K\right) \tag{22.32}$$

where $\mathbf{V} = DD'\sigma_\theta^2 + FF'\sigma_\psi^2 + R$ is the conditional covariance of y implied by (22.25). The REML log-likelihood (i.e., the log-likelihood of $K'y$) is

$$\log L_{\text{REML}} = -\frac{1}{2}(N^* - r)\log 2\pi - \frac{1}{2}\log\left|K'\mathbf{V}K\right| - \frac{1}{2}y'K\left(K'\mathbf{V}K\right)^{-1}K'y. \tag{22.33}$$

The REML estimator of the variance parameters has a number of attractive properties. First, REML estimates are invariant to the choice of K'.[24] Second, REML estimates are invariant to the value of the fixed effects (i.e., β). Third, in the balanced data case, REML is equivalent to ANOVA.[25] Under normality, it thus inherits the minimum variance unbiased property of the ANOVA estimator.[26] Finally, since REML is based on the maximum likelihood principle, it inherits the consistency, efficiency, asymptotic normality, and invariance properties of ML.

Inference based on REML estimates of the variance components parameters is straightforward. Since REML estimation is just maximum likelihood on (22.33), REML likelihood ratio tests (REMLRTs) can be used. In most cases, REMLRTs are equivalent to standard likelihood ratio tests. The exception is testing for the presence of some random effect γ. The null is $\sigma_\gamma^2 = 0$. Denote the restricted REML log-likelihood by $\log L^*_{\text{REML}}$. The REMLRT statistic is $\Lambda = -2\left(\log L^*_{\text{REML}} - \log L_{\text{REML}}\right)$. Since the null puts σ_γ^2 on the boundary of the parameter space under the alternative hypothesis, Λ has a non-standard distribution. Stram and Lee (1994) show the asymptotic distribution of Λ is a 50:50 mixture of a χ_0^2 and χ_1^2. The approximate p-value of the test is thus $0.5\left(1 - \Pr\left(\chi_1^2 \leq \Lambda\right)\right)$.

22.6.2 Estimating the Fixed Effects and Realized Random Effects

A disadvantage of REML estimation is that it provides no direct means for estimating the fixed covariate effects β. Henderson, in Henderson, Kempthorne, Searle and von Krosigk (1959) derived a system of equations that simultaneously yield the BLUE of β and Best Linear Unbiased Predictor (BLUP) of the random effects.

[24] Subject to rows k' satisfying (22.28).

[25] The usual statistical definition of balanced data can be found in Searle (1987). Under this definitions, longitudinal linked data on employers and employees are balanced if we observe each worker employed at every firm, and all job spells have the same duration. Clearly, this is not the usual case.

[26] In contrast, ML estimators of variance components are biased since they do not take into account degrees of freedom used for estimating the fixed effects.

These equations have become known as the mixed model equations or Henderson equations. Define the matrix of variance components

$$G = \begin{bmatrix} \sigma_\theta^2 I_N & 0 \\ 0 & \sigma_\psi^2 I_J \end{bmatrix}. \tag{22.34}$$

The mixed model equations are

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}\begin{bmatrix} D & F \end{bmatrix} \\ \begin{bmatrix} D' \\ F' \end{bmatrix} R^{-1}X & \begin{bmatrix} D' \\ F' \end{bmatrix} R^{-1}\begin{bmatrix} D & F \end{bmatrix} + G^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \tilde{\theta} \\ \tilde{\psi} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ \begin{bmatrix} D' \\ F' \end{bmatrix} R^{-1}y \end{bmatrix} \tag{22.35}$$

where $\tilde{\beta}$ denotes solutions for the fixed effects, and $\tilde{\theta}$ and $\tilde{\psi}$ denote solutions for the random effects. In practice, of course, solving (22.35) requires estimates of R and G. Common practice is to use REML estimates $\tilde{G}$ and $\tilde{R}$.

The BLUPs $\tilde{\theta}$ and $\tilde{\psi}$ have the following properties. They are *best* in the sense of minimizing the mean square error of prediction

$$E\left(\begin{bmatrix} \tilde{\theta} \\ \tilde{\psi} \end{bmatrix} - \begin{bmatrix} \theta \\ \psi \end{bmatrix}\right)' A \left(\begin{bmatrix} \tilde{\theta} \\ \tilde{\psi} \end{bmatrix} - \begin{bmatrix} \theta \\ \psi \end{bmatrix}\right) \tag{22.36}$$

where A is any positive definite symmetric matrix. They are *linear* in y, and *unbiased* in the sense $E(\tilde{\theta}) = E(\theta)$ and $E(\tilde{\psi}) = E(\psi)$.

The solutions to (22.35) also have a Bayesian interpretation. If we suppose that the prior distribution for β is $N(0,\Omega)$ and the prior distribution for (θ,ψ) is $N(0,G)$, then the posterior mean $E[(\beta,\theta,\psi)|y] \to (\tilde{\beta},\tilde{\theta},\tilde{\psi})$, the solution of (22.35), as $|\Omega| \to \infty$. (See Goldberger (1962), Searle et al. (1992, pp. 331–333) and Robinson (1991)).

The mixed model equations make clear the relationship between the fixed and mixed model estimation. In particular, as $|G| \to \infty$ with $R = \sigma_\varepsilon^2 I_{N^*}$, the mixed model equations (22.35) converge to the normal equations (22.18). Thus the mixed model solutions $(\tilde{\beta},\tilde{\theta},\tilde{\psi})$ converge to the least squares solutions $(\hat{\beta},\hat{\theta},\hat{\psi})$. In this sense the least squares estimator is a special case of the mixed model estimator.

22.6.3 Mixed Models and Correlated Random Effects Models

Since Chamberlain (1984) introduced his extension of methods by Cramér (1946) and Mundlak (1978) for handling balanced panel data models with random effects that were correlated with the X variables, econometricians have generally referred to the Chamberlain class of models as "correlated random-effects models." Statisticians, on the other hand, usually mean the Henderson (1953) formulation of the mixed-effects model that gives rise to (22.35), with G nondiagonal, when they refer to a correlated random-effects model.

It is important to distinguish between correlated random-effects models based on the mixed model equations (G nondiagonal) and orthogonal design models, which

can occur within either a fixed-effects or random-effects interpretation of the person and firm effects. Orthogonal design means that one or more of the following conditions hold:

$$X'D = 0, \text{ orthogonal person-effect design and personal characteristics}$$
$$X'F = 0, \text{ orthogonal firm-effect design and personal characteristics}$$
$$D'F = 0, \text{ orthogonal person-effect and firm-effect designs}$$

An economy with random assignment of persons to firms could satisfy these conditions. However, virtually all longitudinal linked employer–employee data, as well as most other observational data in economics, violate at least one of these orthogonal design assumptions. Recognition of the absence of orthogonality between the effects is the basis for the fixed-effects estimator approximations discussed in Sect. 22.5 and the difficulty associated with solving the mixed-model equations, in general (see Robinson, 1991, Searle et al., 1992, Neumaier and Groeneveld, 1996, and Groeneveld, 1998).

To relate the Chamberlain-style correlated random-effects model to the mixed model estimator, we consider a single time-varying X, which we give the components of variance structure:

$$x_{it} = \upsilon_i + \varsigma_{it} \tag{22.37}$$

where

$$\text{Corr}[\upsilon_i, \theta_i] \neq 0$$
$$\text{V}[\varsigma_{it}] = \Delta$$

and

$$\text{Corr}[\varsigma_{it}, \varepsilon_{ns}] = 0 \; \forall i, n, s, t$$

This specification implies that $\text{Corr}[\upsilon_i, \psi_{\text{J}(i,t)}] \neq 0$ as long as G is nondiagonal. Then, to derive the Chamberlain estimating system for a balanced panel data model, assume that $T_i = T$ for all i and compute the linear projection of y_i on x_i

$$y_i = x_i \Pi + \nu_i \tag{22.38}$$

where Π is the $T \times T$ matrix of coefficients from the projection and ν_i is the $T \times 1$ residual of the projection. Chamberlain provides an interpretation of the coefficients in Π that remains valid under our specification.

Because the firm effect is shared by multiple individuals in the sample, however, the techniques proposed by Chamberlain for estimating equation (22.38) require modification. The most direct way to accomplish the extension of Chamberlain's methods is to substitute equation (22.37) into equation (22.7), then restate the system of equations as a mixed model. For each individual i in period t we have

$$\begin{bmatrix} y_{it} \\ x_{it} \end{bmatrix} = \begin{bmatrix} \tau_i + \psi_{\text{J}(i,t)} + \xi_{it} \\ \upsilon_i + \varsigma_{it} \end{bmatrix}. \tag{22.39}$$

where $\tau_i = \theta_i + \upsilon_i\beta$ and $\xi_{it} = \varepsilon_{it} + \varsigma_{it}\beta$. Stacking y_i and x_i, define

$$m_i \equiv \begin{bmatrix} y_i \\ x_i \end{bmatrix}, \text{ and } m \equiv \begin{bmatrix} m_1 \\ \cdots \\ m_N \end{bmatrix}$$

All other vectors are stacked conformably. Then, the mixed-effects formulation of (22.39) can be written as

$$m = D_1\tau + D_2\upsilon + F_3\psi + \nu \tag{22.40}$$

where D_1, D_2, and F_3 are appropriately specified design matrices, τ is the $N \times 1$ vector of person effects entering the y equation, υ is the $N \times 1$ vector of person effects entering the x equation, and

$$\nu = \begin{bmatrix} \xi_1 \\ \varsigma_1 \\ \cdots \\ \xi_N \\ \varsigma_N \end{bmatrix}$$

is the stacked joint error vector. Problems of this form, with τ, υ, and ψ correlated and D_1, D_2, and F_3 nonorthogonal look unusual to economists but are quite common in animal science and statistical genetics. Software to solve the mixed model equations and estimate the variance matrices for (22.40) has been developed by Groeneveld (1998) and Gilmour, Thompson and Cullis (1995) and some applications, other than the one presented above, are discussed in Robinson (1991) and Tanner (1996). The methods exploit the sparse structure of D_1, D_2, and F_3 and use analytic derivatives to solve (22.35). Robert (2001) and Tanner (1996) provide algorithms based on simulated data techniques.

22.7 Models of Heterogeneity Biases in Incomplete Models

The analyses in this section are based upon the exact fixed-effects estimator for model (22.9) given by the solution to (22.18).

22.7.1 Omission of the Firm Effects

When the estimated version of (22.9) excludes the firm effects, ψ, the estimated person effects, θ^*, are the sum of the underlying person effects, θ, and the employment-duration weighted average of the firm effects for the firms in which the worker was employed, conditional on the individual time-varying characteristics, X:

$$\theta^* = \theta + (D'M_X D)^{-1} D'M_X F\psi. \tag{22.41}$$

Hence, if X were orthogonal to D and F, so that $D'M_X D = D'D$ and $D'M_X F = D'F$, then the difference between θ^* and θ, which is just an omitted variable bias, would be an $N \times 1$ vector consisting, for each individual i, of the employment-duration weighted average of the firm effects ψ_j for $j \in \{\mathrm{J}(i,n_{i1}),\ldots,\mathrm{J}(i,n_{iT})\}$:

$$\theta_i^* - \theta_i = \sum_{t=1}^{T_i} \frac{\psi_{\mathrm{J}(i,n_{it})}}{T_i},$$

the person-average firm effect. Similarly, the estimated coefficients on the time-varying characteristics in the case of omitted firm effects, β^*, are the sum of the parameters of the full conditional expectation, β, and an omitted variable bias that depends upon the conditional covariance of X and F, given D:

$$\beta^* = \beta + (X'M_D X)^{-1} X'M_D F\psi.$$

22.7.2 Omission of the Person Effects

Omitting the pure person effects (θ) from the estimated version of (22.9) gives estimates of the firm effects, ψ^{**}, that can be interpreted as the sum of the pure firm effects, ψ, and the employment-duration weighted average of the person effects of all of the firm's employees in the sample, conditional on the time-varying individual characteristics:

$$\psi^{**} = \psi + (F'M_X F)^{-1} F'M_X D\theta. \tag{22.42}$$

Hence, if X were orthogonal to D and F, so that $F'M_X F = F'F$ and $F'M_X D = F'D$, the difference between ψ^{**} and ψ, again an omitted variable bias, would be a $J \times 1$ vector consisting of the employment-duration weighted average of person effects θ_i for $(i,t) \in \{\mathrm{J}(i,t) = j \text{ and } t \in \{n_{i1},\ldots,n_{iT_i}\}\}$ for each firm j. That is,

$$\psi_j^{**} - \psi_j = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \left[\frac{\theta_i 1(\mathrm{J}(i,n_{it}) = j)}{N_j} \right],$$

the firm-average person effect. The estimated coefficients on the time-varying characteristics in the case of omitted individual effects, β^{**}, are the sum of the effects of time-varying personal characteristics in (22.9), β, and an omitted variable bias that depends upon the covariance of X and D, given F:

$$\beta^{**} = \beta + (X'M_F X)^{-1} X'M_F D\theta. \tag{22.43}$$

This interpretation applies to studies like Groshen (1991a, 1991b, 1996).

22.7.3 Inter-industry Wage Differentials

We showed above that industry effects are an aggregation of firm effects that may be inconsistently estimated if either person or firm effects are excluded from the equation. We consider these issues now in the context of inter-industry wage differentials as in Dickens and Katz (1987), Krueger and Summers (1987, 1988), Murphy and Topel (1987), Gibbons and Katz (1992). The fixed or random effects estimation of the aggregation of J firm effects into K industry effects, weighted so as to be representative of individuals, can be accomplished directly by estimation of (22.16). Only $\text{rank}(F'M_{FA}F)$ fixed firm effects can be separately identified; however, the mixed-effects model can produce estimates of all realized industry and firm effects.

As shown in AKM, fixed-effects estimates of industry effects, κ^*, that are computed on the basis of an equation that excludes the remaining firm effects, $M_{FA}F\psi$, are equal to the pure industry effect, κ, plus an omitted variable bias that can be expressed as a function of the conditional variance of the industry effects, FA, given the time-varying characteristics, X, and the person effects, D:

$$\kappa^* = \kappa + \left(A'F'M_{\left[D\ X\right]}FA\right)^{-1} A'F'M_{\left[D\ X\right]}M_{FA}F\psi$$

which simplifies to $\kappa^* = \kappa$ if, and only if, the industry effects, FA, are orthogonal to the subspace $M_{FA}F$, given D and X, which is generally not true even though FA and $M_{FA}F$ are orthogonal by construction. Thus, consistent fixed-effects estimation of the pure inter-industry wage differentials, conditional on time-varying personal characteristics and unobservable non-time-varying personal characteristics requires identifying information on the underlying firms unless this conditional orthogonality condition holds. Mixed-effects estimation without identifying information on both persons and firms likewise produces realized inter-industry wage effects that confound personal and firm heterogeneity.

Similarly, AKM show that fixed-effects estimates of the coefficients of the time-varying personal characteristics, β^*, are equal to the true coefficients of the linear model (22.9), β, plus an omitted variable bias that depends upon the conditional covariance between these characteristics, X, and the residual subspace of the firm effects, $M_{FA}F$, given D:

$$\beta^* = \beta + \left(X'M_{\left[D\ FA\right]}X\right)^{-1} X'M_{\left[D\ FA\right]}M_{FA}F\psi$$

which, once again, simplifies to $\beta^* = \beta$ if, and only if, the time-varying personal characteristics, X, are orthogonal to the subspace $M_{FA}F$, given D and FA, which is also not generally true. Once again, both fixed-effects and mixed-effects estimation of the β coefficients produces estimates that confound personal and firm heterogeneity when both types of identifying information are not available.

To assess the seriousness of the heterogeneity biases in the estimation of industry effects, AKM propose a decomposition of the raw industry effect into the part due to individual heterogeneity and the part due to firm heterogeneity. Their formulas apply directly to the fixed-effects estimator of (22.9) and can be extended to the estimated realized effects in a mixed-effects model. When (22.16) excludes both person and firm effects, the resulting raw industry effect, κ_k^{**}, equals the pure industry effect, κ, plus the employment-duration weighted average residual firm effect inside the industry, given X, and the employment-duration weighted average person effect inside the industry, given the time-varying personal characteristics X:

$$\kappa^{**} = \kappa + (A'F'M_XFA)^{-1}A'F'M_X(M_{FA}F\psi + D\theta)$$

which can be restated as

$$\kappa^{**} = (A'F'M_XFA)^{-1}A'F'M_XF\psi + (A'F'M_XFA)^{-1}A'F'M_XD\theta, \qquad (22.44)$$

which is the sum of the employment-duration weighted average firm effect, given X and the employment-duration weighted average person effect, given X. If industry effects, FA, were orthogonal to time-varying personal characteristics, X, and to the design of the personal heterogeneity, D, so that $A'F'M_XFA = A'F'FA$, $A'F'M_XF = A'F'F$, and $A'F'M_XD = A'F'D$, then, the raw inter-industry wage differentials, κ^{**}, would simply equal the pure inter-industry wage differentials, κ, plus the employment-duration-weighted, industry-average pure person effect, $(A'F'FA)^{-1}A'F'D\theta$, or

$$\kappa_k^{**} = \kappa_k + \sum_{i=1}^{N}\sum_{t=1}^{T_i}\frac{1[\mathrm{K}(\mathrm{J}(i,n_{it})) = k]\theta_i}{N_k}$$

Thus, statistical analyses of inter-industry differentials that exclude either person or firm effects confound the pure inter-industry wage differential with an average of the person effects found in the industry, given the measured personal characteristics, X.

22.8 Endogenous Mobility

The problem of endogenous mobility occurs because of the possibility that individuals and employers are not matched in the labor market on the basis of observable characteristics and the person and firm effects. A complete treatment of this problem is beyond the scope of this article; however, it is worth noting that the interpretation of (22.7) and (22.9) as conditional expectations given the person and firm effects is not affected by some forms of endogenous mobility. If the mobility equation is also conditioned on X, D, and, F, then the effects in the referenced equations are also structural as long as mobility does not depend upon ε.

Matching models of the labor market, such as those proposed by Jovanovic (1979) and Woodcock (2003) imply the existence of a random effect that is the interaction

of person and firm identities. Such models are amenable to the statistical structure laid out in Sect. 22.6; however, to our knowledge the application of such techniques to this type of endogenous mobility model has only been attempted recently using linked employer-employee data. We present these attempts now.

22.8.1 A Generalized Linear Mixed Model

Mixed model theory and estimation techniques have been applied to nonlinear models with linear indices. These are usually called generalized linear mixed models, and include such familiar specifications as the probit, logit, and tobit models augmented to include random effects. See McCulloch and Searle (2001) for a general discussion.

Woodcock (2003) estimates a mixed probit model with random person and firm effects as the first step of a modified Heckman two-step estimator. The goal is to correct for truncation of the error distribution in a mixed model of earnings with random person and firm effects. This truncation arises from endogenous mobility in the context of an equilibrium matching model. Specifically, the Woodcock (2003) matching model predicts that earnings are observed only if the worker-firm match continues, and that the continuation decision depends on person-, firm-, and tenure-specific mobility effects that are correlated with the person and firm effects in the earnings equation. At tenure τ, the match continues only if $\varepsilon_{it} \geq \bar{\varepsilon}_{i\tau}$ where

$$\bar{\varepsilon}_{i\tau} = -\mu_\tau - \zeta_{i\tau} - \xi_{j\tau} \tag{22.45}$$

$$\begin{bmatrix} \zeta_{i\tau} \\ \xi_{j\tau} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\zeta_\tau} & 0 \\ 0 & \sigma^2_{\xi_\tau} \end{bmatrix} \right).$$

When $\varepsilon_{it} \sim \mathcal{N}(0, V_\tau)$, the marginal probability of observing the earnings outcome y_{it} is

$$\begin{aligned} \Pr(\varepsilon_{it} \geq \bar{\varepsilon}_{i\tau}) &= 1 - \Phi\left(\frac{-\mu_\tau - \zeta_{i\tau} - \xi_{j\tau}}{V_\tau^{1/2}} \right) \\ &= \Phi\left(\frac{\mu_\tau + \zeta_{i\tau} + \xi_{j\tau}}{V_\tau^{1/2}} \right) \end{aligned} \tag{22.46}$$

where Φ is the standard normal CDF. Then we have

$$\begin{aligned} E[y_{it} | \varepsilon_{it} \geq \bar{\varepsilon}_{i\tau}] &= \mu + x'_{it}\beta + \theta_i + \psi_j + V_\tau^{1/2} \frac{\phi\left(\frac{\mu_\tau + \zeta_{i\tau} + \xi_{j\tau}}{V_\tau^{1/2}} \right)}{\Phi\left(\frac{\mu_\tau + \zeta_{i\tau} + \xi_{j\tau}}{V_\tau^{1/2}} \right)} \\ &= \mu + x'_{it}\beta + \theta_i + \psi_j + V_\tau^{1/2} \lambda_{i\tau} \end{aligned} \tag{22.47}$$

where $\lambda_{i\tau}$ is the familiar Inverse Mills' Ratio.

The truncation correction based on (22.46) and (22.47) proceeds as follows. The first step is to estimate a continuation probit at each tenure level with random person- and firm-specific mobility effects $\zeta_{i\tau}$ and $\xi_{j\tau}$. Woodcock (2003) estimates probits using the Average Information REML algorithm of Gilmour et al. (1995), applied to the method of Schall (1991). The Schall (1991) method extends standard methods for estimating generalized linear models to the random effects case. The basic idea is to perform REML on a linearization of the link function Φ. The process requires an iterative reweighting of the design matrices of fixed and random effects in the linearized system, see Schall (1991) for details. With estimates of the realized random effects $\tilde{\zeta}_{it}$ and $\tilde{\xi}_{j\tau}$ in hand, Woodcock (2003) constructs an estimate $\tilde{\lambda}_{i\tau}$ of the Inverse Mills' Ratio term for each observation. Including $\tilde{\lambda}_{i\tau}$ as an additional time-varying covariate in the earnings equation corrects for truncation in the error distribution due to endogenous mobility.

22.8.2 A Model of Wages, Endogenous Mobility and Participation with Person and Firm Effects

Following Buchinsky, Fougère, Kramarz and Tchernis (2003), and the structural interpretation they develop, Beffy, Kamionka, Kramarz and Robert (2003, BKKR hereafter) jointly model wages with a participation equation and an inter-firm mobility equation that include state-dependence and unobserved heterogeneity. A firm-specific unobserved heterogeneity component is added to the person-specific term. Like the linear models discussed in detail above, the wage equation includes person and firm effects.

Inter-firm mobility at date t depends on the realized mobility at date $t-1$. Similarly, participation at date t depends on past participation and mobility. Hence, we include initial conditions, modeled following Heckman (1981). This yields the following system of equations:

Initial Conditions:

$$\begin{aligned} z_{i1} &\sim \mathscr{U}_{1,\ldots,J} \\ y_{i1} &= \mathbb{I}\left(X_{i1}^{Y}\delta_0^{Y} + \alpha_{z_{i1}}^{Y,E} + v_{i1} > 0\right) \\ w_{i1} &= y_{i1}\left(X_{i1}^{W}\delta^{W} + \theta_{z_{i1}}^{W,E} + \varepsilon_{i1}\right) \\ m_{i1} &= y_{i1}\mathbb{I}\left(X_{i1}^{M}\delta_0^{M} + \alpha_{z_{i1}}^{M,E} + u_{i1} > 0\right). \end{aligned}$$

Main Equations: $\forall t > 1$,

$$\begin{aligned} z_{it} &= y_{it-1}\left((1-m_{it-1})z_{it-1} + m_{it-1}\widetilde{\eta}_{it}\right) + (1-y_{it-1})\eta \\ \eta &\sim \mathscr{U}_{1,\ldots,J} \quad \widetilde{\eta}_{it} \sim \mathscr{U}_{(1,\ldots,J)-(z_{it-1})} \end{aligned}$$

$$
\begin{aligned}
y_{it} &= \mathbb{I}\left(\underbrace{\gamma^M m_{it-1} + \gamma^Y y_{it-1} + X_{it}^Y \delta^Y + \theta_{z_{it}}^{Y,E} + \theta_i^{Y,I} + v_{it}}_{y_{it}^*} > 0\right) \\
w_{it} &= y_{it}\left(X_{it}^W \delta^W + \theta_{z_{it}}^{W,E} + \theta_i^{W,I} + \varepsilon_{it}\right) \\
m_{it} &= y_{it}\mathbb{I}\left(\underbrace{\gamma m_{it-1} + X_{it}^M \delta^M + \theta_{z_{it}}^{M,E} + \theta_i^{M,I} + u_{it}}_{m_{it}^*} > 0\right).
\end{aligned}
$$

The variable z_{it} denotes the latent identifier of the firm and $J(i,t)$ denotes the realized identifier of the firm at which worker i is employed at date t. Therefore, $J(i,t) = z_{it}$ if individual i participates at date t. y_{it} and m_{it} denote, respectively, participation and mobility, as previously defined. y_{it} is an indicator function, equal to 1 if the individual i participates at date t. m_{it} is an indicator function that takes values according to Table 22.1.

The variable w_{it} denotes the logarithm of the annualized total labor costs. The variables X are the observable time-varying as well as the time-invariant characteristics for individuals at the different dates. Here, θ^I and θ^E denote the random effects specific to, respectively, individuals or firms in each equation. u, v and ε are the error terms. There are J firms and N individuals in the panel of length T.

22.8.3 Stochastic Assumptions

In order to specify the stochastic assumptions for the person and firm-effects, BKKR first rewrite their system of equations as:

$$
\begin{aligned}
z_{it} &= y_{it-1}\left((1 - m_{it-1})z_{it-1} + m_{it-1}\widetilde{\eta}_{it}\right) + (1 - y_{it-1})\eta \\
y_{it} &= \mathbb{I}\left(\underbrace{\gamma^M m_{it-1} + \gamma^Y y_{it-1} + X_{it}^Y \delta^Y + \Omega_{z_{it}}^E \theta^{Y,E} + \Omega_{it}^I \theta^{Y,I} + v_{it}}_{y_{it}^*} > 0\right)
\end{aligned}
$$

Table 22.1 Mobility Indicator

	$y_{it+1} = 1$	$y_{it+1} = 0$
$y_{it} = 1$	$m_{it} = 1$ if $J(i,t+1) \neq J(i,t)$ $m_{it} = 0$ if $J(i,t+1) = J(i,t)$	m_{it} censored
$y_{it} = 0$	$m_{it} = 0$ p.s.	$m_{it} = 0$ p.s.

$$w_{it} = y_{it}\left(X_{it}^{W}\delta^{W} + \Omega_{z_{it}}^{E}\theta^{W,E} + \Omega_{it}^{I}\theta^{W,I} + \varepsilon_{it}\right)$$

$$m_{it} = y_{it}.\mathbb{I}\left(\underbrace{\gamma m_{it-1} + X_{it}^{M}\delta^{M} + \Omega_{z_{it}}^{E}\theta^{M,E} + \Omega_{it}^{I}\theta^{M,I} + u_{it}}_{m_{it}^{*}} > 0\right)$$

for each $t > 1$, where Ω_{it}^{E} is a design matrix of firm effects for the couple (i,t). Hence, it is a $1 \times J$ matrix composed of $J-1$ zeros and of a 1 at column $z_{i,t}$. Similarly, Ω_{it}^{I} is a $1 \times N$ matrix composed of $N-1$ zeros and of a 1 at column i. The model includes two dimensions of heterogeneity. This double dimension crucially affects the statistical structure of the likelihood function. The presence of firm effects makes the likelihood non-separable (person by person). Indeed, the outcomes of two individuals employed at the same firm, not necessarily at the same date, are not independent.

The next equations present the stochastic assumptions for the person and firm effects:

$$\theta^{E} = \left(\alpha^{Y,E}, \alpha^{M,E}, \theta^{Y,E}, \theta^{W,E}, \theta^{M,E}\right) \quad \text{of dimension} \quad [5J, 1]$$
$$\theta^{I} = \left(\theta^{Y,I}, \theta^{W,I}, \theta^{M,I}\right) \quad \text{of dimension} \quad [3N, 1].$$

Moreover,

$$\theta^{E}|\Sigma^{E} \sim \mathcal{N}(0, D_0^{E}) \tag{22.48}$$
$$\theta^{I}|\Sigma^{I} \sim \mathcal{N}(0, D_0^{I}) \tag{22.49}$$
$$D_0^{E} = \Sigma^{E} \otimes I_J \tag{22.50}$$
$$D_0^{I} = \Sigma^{I} \otimes I_N \tag{22.51}$$

where Σ^{E} (resp. Σ^{I}) is a symmetric positive definite matrix $[5,5]$ (resp. $[3,3]$) with mean zero. Notice that these assumptions imply that correlations between the wage, the mobility, and the participation equations come from both person and firm heterogeneity (in addition to that coming from the idiosyncratic error terms). Furthermore, these assumptions exclude explicit correlation between different firms (for instance, the authors could have considered a non-zero correlation of the firm effects within an industry, a non-tractable assumption). Notice though that BKKR could have included in the wage equation, for instance, the lagged firm effects of those firms at which a worker was employed in her career. This is difficult, but feasible in this framework.

Finally, they assume that the idiosyncratic error terms follow:

$$\begin{pmatrix} v_{it} \\ \varepsilon_{it} \\ u_{it} \end{pmatrix} \sim_{iid} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{yw}\sigma & \rho_{ym} \\ \rho_{yw}\sigma & \sigma^2 & \rho_{wm}\sigma \\ \rho_{ym} & \sigma\rho_{wm} & 1 \end{pmatrix}\right).$$

Notice that experience and seniority are complex and highly non linear functions of the participation and mobility equations. Because all these person and firm effects are correlated between equations, the presence of experience and seniority in the wage equation induces a correlation between these two variables and the person and the firm effect in the same equation. Indeed, in the terminology introduced above, the BKKR model exhibits correlated random effects.

BKKR estimate this model on French data using Monte-Carlo Markov Chain methods (Gibbs sampling and the Hastings-Metropolis algorithms).

22.9 Conclusion

We have presented a relatively concise tour of econometric issues surrounding the specification of linear models that form the basis for the analysis of linked longitudinal employer–employee data. Our discussion has focused on the role of person and firm effects in such models, because these data afford analysts the first opportunity to separately distinguish these effects in the context of a wide variety of labor market outcomes. We have shown that identification and estimation strategies depend upon the observed sample of persons and firms (the design of the person and firm effects) as well as on the amount of prior information one imposes on the problem, in particular, the choice of full fixed-effects or mixed-effects estimation.

We do not mean to suggest that these estimation strategies are complete. Indeed, many of the methods described in this chapter have been used by only a few analysts and some have not been used at all in the labor economics context. We believe that future analyses of linked employer–employee data will benefit from our attempt to show the relations among the various techniques and to catalogue the potential biases that arise from ignoring either personal or firm heterogeneity.

References

Abowd, J.M., R. Creecy, and F. Kramarz, "Computing Person and Firm Effects Using Linked Longitudinal Employer–Employee Data," Cornell University working paper, (2002).

Abowd, J.M., H. Finer, and F. Kramarz, "Individual and Firm Heterogeneity in Compensation: An Analysis of Matched Longitudinal Employer and Employee Data for the State of Washington," in J. Haltiwanger, J. Lane, J. Spletzer, K. Troske eds, *The Creation and Analysis of Employer–Employee Matched Data*, Amsterdam, North-Holland, (1999): 3–24.

Abowd, J.M. and F. Kramarz, "The Analysis of Labor Markets Using Matched Employer-Employee Data," in *Handbook of Labor Economics*, O. Ashenfelter and D. Card, (eds). chapter 26, volume 3B, (Amsterdam, North Holland, (1999a): 2629–2710.

Abowd, J.M. and F. Kramarz, "The Analysis of Labor Markets Using Matched Employer–Employee Data," *Labour Economics* 6, (1999b): 53–74.

Abowd, J.M, F. Kramarz, and D.N. Margolis, "High Wage Workers and High Wage Firms," *Econometrica* 67, 2, (1999): 251–333.

Abowd, J.M, F. Kramarz, P. Lengermann, and S. Perez-Duarte, "Are Good Workers Employed by Good Firms? A Test of a Simple Assortative Mating Model for France and the United States," Crest working paper (2003).

Abowd, J., F. Kramarz, P. Lengermann, and S. Roux "Persistent Inter-Industry Wage Differences: Rent-Sharing and Opportunity Costs," Crest Working Paper, (2005).

Abowd, J.M. and M.H. Stinson, "Estimating Measurement Error in SIPP Annual Job Earnings: A Comparison of Census Survey and SSA Administrative Data," US Census Bureau LEHD Program Working Paper (2003).

Abowd, J.M. and S. Woodcock, "Disclosure Limitation in Longitudinal Linked Data," in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes (eds.), Amsterdam: North Holland, (2001), 215–277.

Buchinsky M., D. Fougère, F. Kramarz, and R. Tchernis, "Interfirm Mobility, Wages, and the Returns to Seniority and Experience in the U.S.," Crest working paper (2003).

Beffy M., T. Kamionka, F. Kramarz, and C.P. Robert, "Job Mobility and Wages with Worker and Firm Heterogeneity", Crest working paper, (2003).

Chamberlain, G., "Panel Data," in *Handbook of Econometrics*, ed. by Z. Griliches and M.D. Intrilligator. Amsterdam: North Holland, chapter 22, (1984): 1248–1318.

Cramér, H., *Mathematical Models of Statistics* (Princeton, NJ: Princeton University Press, (1946).

Dickens, W.T. and L.F. Katz, "Inter-Industry Wage Differences and Industry Characteristics," in *Unemployment and the Structure of Labor Markets*, (eds.) Kevin Lang and Jonathan S. Leonard (eds.). Oxford: Basil Blackwell (1987).

Dongarra, J., I. Duff, D. Sorensen and H. Van der Vorst, *Solving Linear Systems on Vector and Shared Memory Computers,* Philadelphia: SIAM, (1991).

Dostie, B., "Job Turnover and the Returns to Seniority," *Journal of Business and Economic Statistics,* 23, 2, (2005): 192–199.

Gibbons, R. and L. Katz, "Does Unmeasured Ability Explain Inter-Industry Wage Differentials?" *Review of Economic Studies*, 59, (1992): 515–535.

Gilmour, A.R., R. Thompson, and B.R. Cullis, "Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models." *Biometrics*, 51, (1995): 1440–1450.

Goldberger, A.S., "Best Linear Unbiased Prediction in the Generalized Linear Regression Model," *Journal of the American Statistical Association*, 57 (1962): 369–75.

Groeneveld, E., *VCE4 User's Guide and Reference Manual* (Höltystrass, Germany: Institute of Animal Husbandry and Animal Behavior, (1998).

Groshen, E., "Sources of Intra-Industry Wage Dispersion: How Much do Employers Matter?" *Quarterly Journal of Economics,* 106, (1991a): 869–884.

Groshen, E., "The Structure of the Female/Male Wage Differential: Is it Who You Are, What You Do, or Where You Work?" *Journal of Human Resources*, 26, (1991b): 457–472.

Groshen, E., "American Employer Salary Surveys and Labor Economics Research: Issues and Contributions," *Annales d'économie et de statistique*, 41/42, (1996): 413–442.

Heckman J.J., "Heterogeneity and State Dependence", in *Studies in Labor Market*, Rosen S. (ed.) University of Chicago Press, (1981).

Henderson, Charles Roy, "Estimation of Variance and Covariance Components," *Biometrics* 9 (1953): 226–252.

Henderson, Charles Roy, O. Kempthorne, S.R. Searle and C.M. Von Krosigk, "The Estimation of Environmental and Genetic Trends from Records Subject to Culling," *Biometrics*, 15, 2, (1959): 192–218.

Jovanovic, B.,"Job Matching and the Theory of Turnover," *Journal of Political Economy,* 87, (1979): 972–990.

Krueger, A.B. and L.H. Summers, "Reflections on the Inter-industry Wage Structure," in *Unemployment and the Structure of Labor Markets*, Kevin Lang and Jonathan S. Leonard (eds.), New York: Basil Blackwell, (1987).

Krueger, A.B. and L.H. Summers, "Efficiency Wages and the Inter-Industry Wage Structure," *Econometrica*, 56, (1988): 259–293.

Lane, J., S. Burgess and J. Theeuwes, "The Uses of Longitudinal Matched Employer/Employee Data in Labor Market Analysis." Proceedings of the American Statistical Association, (1998).

Lillard, L.A., "Job Turnover Heterogeneity and Person-Job-Specific Time-Series Wages," *Annales D'Économie et de Statistique*, 55–56 (1999): 183–210.

Lillard, L.A. and C.W. Panis, *aML Multilevel Multiprocess Statistical Soft-ware*, Release 1.0 EconWare, Los Angeles, California (2000).

McCulloch C.E. and S.R. Searle, *Generalized, Linear, and Mixed Models*, New York: John Wiley and Sons (2001).

Mundlak, Y., "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46 (1978): 69–85.

Murphy, K.M. and R.H. Topel, "Unemployment, Risk, and Earnings: Testing for Equalizing Wage Differences in the Labor Market" in *Unemployment and the Structure of the Labor Market*, Kevin Lang and Jonathan S. Leonard (eds.). New York: Basil Blackwell, (1987).

Neumaier, A. and E. Groeneveld, "Restricted Maximum Likelihood Estimation of Covariance in Sparse Linear Models," working paper, Institut für Mathematik, Wien University, Austria, (1996).

Raghunathan, T.E., J.M. Lepkowski, J. Van Hoewyk, and P. Solenberger, "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 27, (2001): 85–95.

Robert C.P., *The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation*, New York: Springer-Verlag, (2001).

Robinson, G.K., "That BLUP is a Good Thing: The Estimation of Random Effects," *Statistical Science*, 6, (1991): 15–51.

Schall R., "Estimation in Generalized Linear Models with Random Effects," *Biometrika*, 78, 4, (1991): 719–727.

Searle, S.R., *Linear Models for Unbalanced Data*, New York: John Wiley and Sons, 1987.

Searle, S.R., George Casella and Charles E. McCulloch, *Variance Components* New York: John Wiley and sons, (1992).

Stram, D.O. and J.W. Lee, "Variance Component Testing in the Longitudinal Mixed Effects Model," *Biometrics*, 50, (1994): 1171–1177.

Tanner, M.A., *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* New York: Springer, (1996).

Weeks, D.L. and D.R. Williams, "A Note on the Determination of Connectedness in an *N*-way cross classification." *Technometrics*, 6, (1964): 319–324.

Woodcock, S.D., "Heterogeneity and Learning in Labor Markets", in *Essays on Labor Market Dynamics and Longitudinal Linked Data*, Cornell University Ph.D. Thesis (2003).

Chapter 23
Life Cycle Labor Supply and Panel Data: A Survey

Bertrand Koebel, François Laisney, Winfried Pohlmeier and Matthias Staat

23.1 Introduction

The econometrics of labor supply belongs to one of the technically most advanced fields in microeconometrics. Many specific issues such as the proper modelling of tax structures, the existence of fixed costs as well as rationing have been treated in numerous articles so that marginal gains in substantive economic insights seem low and entry costs into the field prohibitively high. Not surprisingly, one of the most obvious paths for research on labor supply, the (micro-) econometric analysis of the individual's labor supply over the life cycle, has by now gained much more attention than 10 years ago. The increased availability of panel data for many countries, as well as the development of appropriate econometric techniques, have made econometric studies of intertemporal labor supply behavior using panel data not only interesting on purely theoretical grounds, they have also helped to achieve a better understanding of individual retirement behavior, the functioning of institutional settings in different countries (such as taxes, vocational training programmes, day-care for children) and the distribution of income and wealth, to name only a few.

Estimation of labor supply functions using panel data has started out in the eighties, and the number of studies reporting on such estimation is rapidly increasing.

Bertrand Koebel
BETA, Université Louis Pasteur, Strasbourg I and IZA, Bonn, 61 Avenue de la Forêt Noire, F67000 Strasbourg, France, e-mail: koebel@cournot.u-strasbg.fr

François Laisney
BETA, Université Louis Pasteur, Strasbourg I and ZEW, Mannheim, 61 Avenue de la Forêt Noire, F67000 Strasbourg, France, e-mail: fla@cournot.u-strasbg.fr

Winfried Pohlmeier
University of Konstanz, Department of Economics, Box D124, D78457 Konstanz, Germany, e-mail: winfried.pohlmeier@uni-konstanz.de

Matthias Staat
University of Mannheim, Department of Economics, D68131 Mannheim, Germany, e-mail: matthias@pool.uni-mannheim.de

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

Earlier studies using panel data mainly concentrated on participation. Thus, it is not surprising that the excellent surveys of Pencavel (1986), Heckman and MaCurdy (1986) and Killingsworth and Heckman (1986) hardly touched the subject.[1] The latter survey concluded a comparison of a large number of cross section studies with the words: "[these studies] seem to have reduced the mean and substantially increased the variance of [...] what might be called the *reasonable guesstimate* of the wage elasticity of female labour supply [...].[2] However, [...] studies based on alternative behavioural models—notably, life cycle models, which have been used relatively little in empirical studies—are also likely to provide important insights" (pp. 196–197).

Earlier surveys of some of the material covered here can be found in Blundell (1987, 1988), Blundell, Fry and Meghir (1990), Card (1994), MaCurdy et al. (1990) and Blundell and MaCurdy (1999).

As we shall see, there has been a trend away from models that take advantage of panel data almost exclusively in order to control for unobserved heterogeneity, towards fully dynamic models where wages become endogenous, and consequently the concept of wage elasticity loses much of its appeal.

This chapter aims at providing the reader with a thread through the literature on the topic. However, we make no claim to exhaustivity, and concentrate mainly on the theoretical aspects of the studies. In Sect. 23.2 we describe the basic model of life cycle λ-constant labor supply. Sect. 23.3 is devoted to extensions taking account of uncertainty and risk, while Sect. 23.4 discusses voluntary and involuntary non-participation, as well as accounting for taxation. Sect. 23.5 presents an alternative specification which leaves the λ-constant framework, and discusses its implications, in particular for modelling the impact of taxes on labor supply. In Sect. 23.6 we discuss studies relaxing within-period and between-period additive separability, and focusing on rational habit formation and human capital accumulation. Sect. 23.7 concludes and opens towards other strands of the literature that contribute to the understanding of labor supply.

23.2 The Basic Model of Life Cycle Labor Supply

We shall not restate here the theoretical developments contained in the survey of Killingsworth and Heckman (1986) (pp. 144–179) but refer the reader to them. Killingsworth and Heckman insist on the pioneering work of Mincer (1962). They show that "the distinction between permanent and transitory wages is not particularly useful from a theoretical standpoint" (p. 158) and demonstrate the usefulness of Frisch demands as an alternative to the permanent vs. transitory distinction.[3] They also discuss models with endogenous wages and conclude: "although much informal discussion implicitly or explicitly emphasizes the interrelationships be-

[1] Yet see Chap. 5 in Killingsworth (1983), pp. 207–330.

[2] Here we shall not restrict attention to female labor supply.

[3] The uninformed reader will find a definition below.

tween (...) work and wages in a life-cycle setting, rigorous analysis of such issues using formal life-cycle labour supply models with endogenous wages is still in its infancy" (p. 178). Here we will describe the models used for estimation in a selection of papers representative of the trend over the last 25 years. Along the way we also give some details on the estimation techniques and on the results, illustrating the fact that econometric modelling is by no means linear: there is a feedback of estimation results on model specification.

23.2.1 The Framework

The seminal paper, as far as empirically implementable models are concerned, is MaCurdy (1981).[4] The assumptions retained are fairly stringent and include known life length T, perfect foresight and perfect credit markets, as well as rates of time preference that may differ across individuals and do not change over time. At time 1 an individual chooses $\{C_{it}, L_{it}, A_{it}\}_{t=1}^{T}$ in order to maximizes discounted utility

$$\sum_{t=1}^{T} \frac{1}{(1+\rho_i)^{t-1}} U_{it}(C_{it}, L_{it}) \tag{23.1}$$

subject to the sequence of budget constraints

$$A_{it} = (1+r_t)A_{i,t-1} + w_{it}N_{it} - C_{it}\,, \quad t = 1,\ldots,T\,. \tag{23.2}$$

The variable C denotes real consumption, L leisure, A end of period assets in real terms, N hours of work ($N = \bar{L} - L$, where $\bar{L}$ denotes maximum time available in each period for allocation between leisure and market work), r is the real interest rate, w the real wage, ρ the rate of time preference, and A_0 denotes initial assets. The within-period utility function U_{it} is assumed to be concave.

The first-order conditions, assuming an interior optimum, include the budget restrictions (23.2) and

$$\frac{\partial U_{it}}{\partial C_{it}} = \lambda_{it}, \tag{23.3}$$

$$\frac{\partial U_{it}}{\partial L_{it}} = \lambda_{it} w_{it}, \quad t = 1,\ldots,T, \tag{23.4}$$

where λ_{it} denotes the Lagrange multiplier of the budget constraint in period t. Notice that (pseudo) optimal demands can be derived by solving (23.2)–(23.4), eliminating λ_{it}, to obtain $\widetilde{C}_{it}(S_{it}, w_{it})$, $\widetilde{L}_{it}(S_{it}, w_{it})$, where $S_{it} \equiv A_{it} - (1+r_t)A_{i,t-1}$ denotes the level of saving or dissaving.[5] In the timewise additive separable case, net saving S_{it} is a sufficient statistic of all the future as far as the present decision is concerned. In

[4] For the purpose of comparability with later sections, we slightly depart from MaCurdy's exposition and notations.

[5] MaCurdy (1983, p.271) calls $\widetilde{C}$ and $\widetilde{L}$ pseudo demand functions.

general, the argument S_{it} of functions $\widetilde{C}_{it}$ and $\widetilde{L}_{it}$ will not be arbitrary, but optimally chosen by individuals. For this reason it will depend of the entire wage profile, of the initial wealth A_{i0} and of the interest and time preference rates. This functional dependence in general implies correlation between S_{it} and the past and future variables, and shocks and thus calls for instrumental variable estimation methods.

Instead of considering $\widetilde{C}$ and $\widetilde{L}$, MaCurdy (1981) derives the Frisch demands $C_{it}(\lambda_{it}, w_{it})$, $L_{it}(\lambda_{it}, w_{it})$, obtained by solving (23.3)–(23.4). The Lagrange multiplier λ_{it} measures the impact of a marginal increase in A_{it} on the optimal value of objective (23.1). From the envelope theorem, we have

$$\lambda_{it} = \frac{1 + r_{t+1}}{1 + \rho_i} \lambda_{it+1} , \tag{23.5}$$

or, using a first-order approximation around $\rho_i = r_{t+1} = 0$,

$$\ln \lambda_{it} \approx r_{t+1} - \rho_i + \ln \lambda_{i,t+1} . \tag{23.6}$$

The value of λ_{it} is implicitly determined by substitution of the demand functions C and L in (23.2). Thus, λ_{it} is a function of the entire wage profile, of the initial wealth A_{i0} and of the interest and time preference rates r_t and ρ_i. Just as S_{it}, λ_{it} is a sufficient statistic which summarizes the impact of all the future variables on the present decision. As before, the use of instrumental variables is recommended for parameter estimation. Using (23.5) and (23.6), we can write

$$\lambda_{it} = \lambda_{i0} \prod_{k=1}^{t} \frac{1 + \rho_i}{1 + r_k} \tag{23.7}$$

or, assuming small values for ρ_i and the r_k,

$$\ln \lambda_{it} \approx t\rho_i - R_t + \ln \lambda_{i0} , \tag{23.8}$$

where $R_t = \sum_{k=1}^{t} r_k$, and substitute this term into functions C and L to obtain

$$C^*_{it}(w_{it}, \lambda_{i0}) \quad \text{and} \quad L^*_{it}(w_{it}, \lambda_{i0}) .$$

The concavity of U_{it} implies

$$\begin{aligned} &\frac{\partial C^*_{it}}{\partial w_{it}} \geq 0, \quad \frac{\partial L^*_{it}}{\partial w_{it}} \leq 0, \\ &\frac{\partial C^*_{it}}{\partial \lambda_{i0}} \leq 0, \quad \frac{\partial L^*_{it}}{\partial \lambda_{i0}} \leq 0, \quad \frac{\partial^2 L^*_{it}}{\partial \lambda^2_{i0}} \leq 0, \\ \text{and} \quad &\frac{\partial \lambda^*_{i0}}{\partial A_{i0}} \leq 0, \quad \frac{\partial \lambda^*_{i0}}{\partial w_{it}} \leq 0, \quad t = 1, \ldots, T. \end{aligned} \tag{23.9}$$

where $\lambda^*_{i0}\left(\{w_{it}\}_{t=0}^{T}, A_{i0}\right)$ is the value of the multiplier corresponding to the optimal solution.

Both types of demand functions are related by:

$$L^*(\lambda_{i0}^*, w_{it}) = \widetilde{L}(S_{i0}, w_{it}) \ .$$

Two measures have focused the interest of economists: the Frisch elasticity of labor supply with respect to the wage, denoting $N_{it}^* = N^*(\lambda_{i0}, w_{it})$,

$$e_\lambda \equiv \frac{\partial N_{it}^*}{\partial w_{it}} \frac{w_{it}}{N_{it}^*} \ ,$$

and the intertemporal elasticity of substitution between labor supplies of two consecutive periods:

$$ies \equiv \frac{\partial \left(N_{it}^* / N_{i,t+1}^* \right)}{\partial \left(w_{it} / w_{i,t+1} \right)} \frac{w_{it} / w_{i,t+1}}{N_{it}^* / N_{i,t+1}} \ ,$$

which gives the inverse of the percentage change in the relative labor supplies (of two consecutive periods), when the ratio of relative wages $w_{it}/w_{i,t+1}$ increases by 1%.

23.2.2 First Specifications of the Utility Function

MaCurdy (1981) specifies the following additively separable within-period utility function for individual i:

$$U_{it}(C_{it}, L_{it}) = \gamma_{Cit} C_{it}^{\beta} - \gamma_{Nit} N_{it}^{\alpha_N}, \quad [N_{it} = \bar{L} - L_{it}], \quad i = 1, \ldots, I \ . \tag{23.10}$$

Concavity requires $0 < \beta < 1$, $\alpha_N > 1$. Heterogeneity, both observed and unobserved, is modelled through random preferences with the specification

$$\ln \gamma_{Nit} = \sigma_i - u_{it}^* \ , \tag{23.11}$$

where u_{it}^* is i.i.d. with zero expectation (note that time–varying characteristics are excluded by assumption).

The resulting Frisch labor supply and consumption demand equations are:

$$\begin{aligned} \ln N_{it} &= \frac{1}{\alpha_N - 1} (\ln \lambda_{it} - \ln \alpha_N + \ln w_{it} - \sigma_i + u_{it}^*) \\ \ln C_{it} &= \frac{1}{\beta - 1} (\ln \lambda_{it} - \ln \gamma_{Cit} - \ln \beta). \end{aligned} \tag{23.12}$$

Using (23.8), we obtain (assuming $\rho_i = \rho$)

$$\ln N_{it} = F_i + bt - \delta R_t + \delta \ln w_{it} + u_{it} \tag{23.13}$$

with

$$F_i = \frac{1}{\alpha_N - 1}(\ln \lambda_{i0} - \sigma_i - \ln \alpha_N), \quad \delta = \frac{1}{\alpha_N - 1}, \quad b = \delta\rho, \quad u_{it} = -\delta u_{it}^* .$$

This is a linear panel model with an individual-specific effect F_i, which has to be treated as a fixed effect because it is correlated with w_{it} via λ_{i0}. Notice that when the ρ_i are not all identical, there is in addition heterogeneity in the parameter b. In this model, the Frisch elasticity of labor supply is given by $\delta = 1/(\alpha_N - 1)$ and is also equal to the intertemporal elasticity of substitution.

Moreover, MaCurdy considers the following linear approximation of F_i:

$$F_i = Z_i\phi + \sum_{t=1}^{T} \gamma_t \ln w_{it} + A_{i0}\theta + \alpha_i , \tag{23.14}$$

where Z_i denotes a vector of household characteristics and α_i a residual term. According to (23.9), γ_t and θ should be negative. Note that coefficients are identical across households. Combined with the additional assumption of a quadratic form for the profile of log wages,

$$\ln w_{it} = \pi_{0i} + \pi_{1i}t + \pi_{2i}t^2 + \xi_{it} , \tag{23.15}$$

this leads to

$$F_i = Z_i\phi + \pi_{0i}\overline{\gamma}_0 + \pi_{1i}\overline{\gamma}_1 + \pi_{2i}\overline{\gamma}_2 + A_{i0}\theta + \eta_i , \tag{23.16}$$

with

$$\overline{\gamma}_j = \sum_{t=1}^{T} \gamma_t t^j, \qquad j = 0, 1, 2 .$$

Interpretation: δ is the intertemporal substitution (or λ-constant, or Frisch) elasticity. It describes the reaction to an evolutionary change of the wage rate along the wage profile. It is positive since $\alpha_N > 1$. *Along* a profile, *evolutionary* changes take place. MaCurdy calls changes *between* profiles *parametric* or profile changes. A change $\Delta > 0$ from a wage profile I to an otherwise identical profile II at time s causes the Frisch labor supply of profile II to be lower than that of profile I in all periods $t \neq s$, because $\lambda_{II} < \lambda_I$ by (23.9). Equation (23.14) implies

$$F_{II} - F_I = \gamma_s\Delta < 0 .$$

The net effect on labor supply in period s, $(\delta + \gamma_s)\Delta$, can be positive or negative. $\delta + \gamma_s$ and γ_s are the usual uncompensated (own- and cross-period) elasticities, and the corresponding compensated elasticities are $\delta + \gamma_s - \mathrm{E}_s\theta$ and $\gamma_s - \mathrm{E}_s\theta$, respectively, where E_s denotes real earnings in period s. If leisure is a normal good ($\theta < 0$), we have

$$\delta > \delta + \gamma_s - \mathrm{E}_s\theta > \delta + \gamma_s ,$$

i.e.

$$e_\lambda > e_u > e_A ,$$

where e_λ is the wage elasticity with constant marginal utility of wealth, e_A is the wage elasticity with constant (lifetime) wealth and e_u is the wage elasticity with constant (lifetime) utility. Bover (1989) and Blundell, Meghir and Neves (1993) give useful discussions of the relationships between these elasticities.

Estimation is conducted in two stages.

Stage 1: (23.12) is estimated in first differences:[6]

$$\Delta \ln N_{it} = b - \delta r_{t+1} + \delta \Delta \ln w_{it} + \varepsilon_{it}, \quad t = 2, \ldots, \tau, \quad i = 1, \ldots, I \, .$$

MaCurdy (1981) considers the Frisch labor supply equations across the τ available time periods as a system. No restrictions are imposed on the temporal covariance structure of ε. As the level of wages may depend upon unobserved individual characteristics which also affect the amount of working time, the variable w_{it} can be suspected to be correlated with ε_{it}. MaCurdy uses system estimation (2SLS and 3SLS), and treats $\ln w_{it}$ as endogenous, with instruments derived from a human capital type equation.

In this way, the reactions of N_{it} to the *evolutionary* changes in w_{it} are completely described by $\hat{\delta}$. In order to also describe the reactions of labor supply to *parametric* changes in wages, information on the sensitivity of F_i with respect to w_{it} is needed.

Stage 2: Given the first stage parameter estimates, the fixed effects can be estimated using (23.12) as:

$$\hat{F}_i = \frac{1}{\tau} \sum_{t=1}^{\tau} \left(\ln N_{it} - \hat{b}t + \hat{\delta} R_t - \hat{\delta} \ln w_{it} \right) \, . \tag{23.17}$$

A similar method is used to obtain estimates of the π_{hi} parameters, which then allows to estimate the unknown parameter of (23.16). These estimates can then be used to identify the labor demand reaction to a shift in the wage profile and to obtain an estimate of the wage elasticity e_A.

Note that there are also contributions estimating pseudo supply functions. For instance, Conway and Kniesner (1994) consider the following econometric specification:

$$N_{it} = F_i + \delta w_{it} + \kappa S_{it} + Z_{it} \pi + u_{it} \, ,$$

which is a linear pseudo labor function (depending upon savings), where variables w_{it} and/or S_{it} are allowed to be correlated with the random term u_{it} and individual specific heterogeneity F_i. They use a sample of prime aged men from the PSID who worked each year from 1978 to 1982 and experiment with different types of instruments. They find that pseudo labor supply is decreasing in the wage in 59 out of the 60 regressions considered. This finding is at odds with (23.9).

[6] Henceforth, Δ will denote the first difference operator. Another possibility would be to use within estimation. One advantage of estimation in first differences, however, is that no strict exogeneity assumption is needed.

23.3 Taking Account of Uncertainty and Risk

So far we have only considered labor substitution over time, which measures changes in labor supply in response to *anticipated* wage changes. How individuals react in response to *unanticipated* wage and interest rate changes is important to better understand the labor market impacts of monetary and fiscal policies for example. The labor market implications of wage and interest rate volatility may also have consequences for the optimal design of labor contracts and the organization of financial markets.

MaCurdy (1983) was the first to propose an empirical framework allowing to cope with uncertainty. He showed that uncertainty concerning wages and interest rates can be accounted for by slightly adapting the model with certainty, so that most uncertainty can be summarized into an additive residual term. The use of adequate instruments then allows to consistently estimate the parameters of interest. Some 20 years later, Pistaferri (2003) showed that a more precise modelling of uncertainty yields a different specification of labor supply relationships. This allows economists to study how labor supply reacts to unanticipated changes not only in wages and interest rates, but also in other dimensions like wealth or family composition. As soon as uncertainty is introduced in the model, risk also naturally arises in the specification of labor supply. Lich-Tyler (2002) investigated this second issue. Both topics are related and can be presented within a comprehensive framework.

23.3.1 First Developments

Following MaCurdy (1983), we assume uncertainty concerning future wages and interest rates. Replanning for the future takes place in every period, on the basis of the new information obtained. The individual maximizes expected discounted utility in period t:

$$\mathrm{E}_t \sum_{s=t}^{T} \frac{1}{(1+\rho_i)^{s-t}} U_{is}(C_{is}, L_{is}) \,, \tag{23.18}$$

subject to the budget restriction (23.2). If we exclude corner solutions, the first-order conditions include (23.3) and (23.4) at period $t = 1$.

As in static models, the ratio of first derivatives is still equal to relative prices, so that this can provide the basis for estimating demand elasticities. This estimation strategy was followed by MaCurdy (1983), using instrumental variables for controlling the endogeneity of C_{it} and L_{it}. Note that also the functions C_{it}^* and L_{it}^* are just the same as in the certain case. This might suggest that differences between the certain and uncertain cases are not important in the time additive separable case. However, the level of saving S_{it} chosen in period t for some configuration of expected future wage and interest rate paths, can turn out not having been optimal *ex post*, once time discloses additional information. This is why replanning is necessary at each period.

The Lagrange multipliers now satisfy

$$\lambda_{it} = \mathrm{E}_t \left(\frac{1+r_{t+1}}{1+\rho_i} \lambda_{i,t+1} \right) , \tag{23.19}$$

implying that the individual decides on savings in such a way that the discounted expected utility of wealth remains constant. If we assume that there is no uncertainty about r_{t+1} we have

$$\lambda_{it} = \frac{1+r_{t+1}}{1+\rho_i} \mathrm{E}_t \lambda_{i,t+1} ,$$

which leads to the (first-order) *approximation*

$$\begin{aligned} \ln \lambda_{it} &\approx \mathrm{E}_t \ln \lambda_{i,t+1} - \rho_i + r_{t+1} \\ &= \ln \lambda_{i,t+1} - \rho_i + r_{t+1} + e_{i,t+1}, \end{aligned} \tag{23.20}$$

where the random term $e_{i,t+1}$, a forecast error of the marginal utility of next period, satisfies $\mathrm{E}_t(e_{i,t+1}) = 0$. Once substituted in the λ-constant demands in first difference obtained from (23.12):

$$\Delta \ln N_{it} \approx \delta \Delta \ln w_{it} + \delta \left(\ln \lambda_{i,t+1} - \ln \lambda_{it} \right) , \tag{23.21}$$

this yields

$$\Delta \ln N_{it} \approx \delta \Delta \ln w_{it} + \delta \left(\rho_i - r_{t+1} \right) - \delta e_{t+1} . \tag{23.22}$$

From (23.19) and the expression $N(\lambda_{it}, w_{it})$ of Frisch labor supply, it can be seen that expected changes in $\lambda_{i,t+1}$ are already taken into account for determining labor supply at period t. As a consequence, only unexpected changes in the marginal utility of wealth influence changes in labor supply through $e_{i,t+1}$. This is the economic interpretation of the residual term in (23.22).

As in the certain case, the λ-constant demands can be relied on for estimation. The "fixed effects" techniques remain available in the presence of uncertainty about the wage profile. Under rational expectations, the orthogonality between $e_{i,t+1}$ and the information available at time t suggests application of the Generalized Method of Moments (GMM). Exposition here has been kept fairly sketchy, and we refer the reader to Altug and Miller (1990) for a more elaborate treatment spelling out the implications of assuming a competitive environment with complete markets.

Others contributions in this vein investigate the impact of unexpected capital, windfall gains, house price shocks and inheritance on labor supply: see Joulfaian and Wilhelm (1994) and Henley (2004). Both studies report that unexpected gains exert (mostly) significant negative effects on working hours, but their impact is relatively small in absolute value.

23.3.2 Recent Contributions

Now we turn to the contributions of Pistaferri (2003) and Lich-Tyler (2002) , who derive a labor supply specification from a more precise approximation of the relationship between consecutive marginal utilities of wealth (23.19). Although we do not follow exactly each author's presentation, we hope that our interpretation does a good job of summarizing the main novelty of both contributions.

Without (intra-period) additive separability between consumption and leisure (see next section), the λ-constant demands in first differences (23.21) become

$$\Delta \ln N_{it} \approx \delta \Delta \ln w_{it} + \eta \left(\ln \lambda_{i,t+1} - \ln \lambda_{it} \right) , \tag{23.23}$$

where $\eta > \delta$ when C and L are substitutes and $\eta < \delta$ when they are complements. Instead of approximating $\ln \lambda_{it}$ by (23.20), let us use a second order Taylor approximation to the random function $\lambda_{t+1}(1+r_{t+1})/(1+\rho)$ in the neighborhood of its arguments' mean and take its expectation to obtain

$$\begin{aligned} \ln \lambda_{it} &\approx \ln \left[\frac{1+\mathrm{E}_t r_{t+1}}{1+\rho} \mathrm{E}_t \lambda_{i,t+1} \right] \\ &+ \frac{1}{2(1+\rho)} \mathrm{E}_t \begin{pmatrix} r_{t+1} - \mathrm{E}_t (r_{t+1}) \\ \lambda_{i,t+1} - \mathrm{E}_t (\lambda_{i,t+1}) \end{pmatrix}' \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} r_{t+1} - \mathrm{E}(r_{t+1}) \\ \lambda_{i,t+1} - \mathrm{E}(\lambda_{i,t+1}) \end{pmatrix} \\ &\approx \ln \mathrm{E}_t \lambda_{i,t+1} - \rho + \mathrm{E}_t r_{t+1} + \frac{\mathrm{Cov}_t (r_{t+1}, \lambda_{i,t+1})}{(1+\rho)} . \end{aligned} \tag{23.24}$$

Similarly, it can be shown that:[7]

$$\mathrm{E}_t \ln \lambda_{i,t+1} \approx \ln \mathrm{E}_t \lambda_{i,t+1} - \frac{\mathrm{Var}_t (\lambda_{i,t+1})}{2 (\mathrm{E}_t \lambda_{i,t+1})^2} .$$

Replacing these expressions into (23.23) yields

$$\begin{aligned} \Delta \ln N_{it} &\approx \eta (\rho - r_{t+1}) + \delta \Delta \ln w_{it} + \eta \left(r_{t+1} - \mathrm{E}_t r_{t+1} - \frac{1}{1+\rho} \mathrm{Cov}_t (r_{t+1}, \lambda_{i,t+1}) \right) \\ &+ \eta \left(\ln \lambda_{i,t+1} - \mathrm{E}_t \ln \lambda_{i,t+1} - \frac{\mathrm{Var}_t (\lambda_{i,t+1})}{2 (\mathrm{E}_t \lambda_{i,t+1})^2} \right) . \end{aligned} \tag{23.25}$$

This is the extended λ-constant labor supply relationship which depends on two new kinds of explanatory variables: (i) innovations in the marginal utility of wealth and interest rate, and (ii) risk in the marginal utility of wealth and interest rate, reflected in the variance–covariance terms. An increase in $\mathrm{Var}_t (\lambda_{i,t+1})$ has the same effect as reducing the marginal utility of wealth at period $t+1$. From economic theory, we

[7] For any positive random variable, say v, it can be seen that $\mathrm{E}_t \ln v \approx \ln \mathrm{E}_t v - \mathrm{V}_t v \,/\, \left[2 (\mathrm{E}_t v)^2 \right]$.

expect that $\eta > 0$, which means that a greater than expected interest rate increases current labor supply. Similarly, individuals or time periods with high risk (reflected by the variance terms) are characterized by a more decreasing labor supply profile than individuals/periods with low risk.[8]

In order to obtain an empirically tractable expression for labor supply dynamics, it is necessary to find an observable analogue for the last terms in (23.25). Hence, it is necessary to understand how the marginal utility of wealth evolves over the life cycle. Two strategies have been relied on for this purpose. Pistaferri (2003) translates the uncertainty and risk on marginal utility of wealth $\lambda_{i,t+1}$ into uncertainty and risk on wages. His strategy relies on two assumptions; one about the expectation error (assumed to follow an MA(1) process), and one linking the marginal utility of wealth to wages, as in (23.14). Lich-Tyler (2002) relies on definition of the marginal utility of wealth to obtain an estimable expression for $\ln \lambda_{i,t+1} - \mathrm{E}_t \ln \lambda_{i,t+1}$.

We follow Lich-Tyler's strategy to derive a simplified version of the model. In our time separable framework, let us define the period t indirect utility function:[9]

$$V(r_t, w_{it}, A_{it}) = \max_{C,N} \left\{ U(C, \overline{N} - N) : (1 + r_t) A_{i,t-1} + w_{it} N = C + A_{it} \right\} .$$

Then

$$\lambda_{it} = \frac{\partial V}{\partial A}(r_t, w_{it}, A_{it}) , \tag{23.26}$$

which can be used to obtain an expression for $\ln \lambda_{i,t+1} - \mathrm{E}_t \ln \lambda_{i,t+1}$. Using a first-order Taylor approximation to

$$\lambda_{i,t+1} = \frac{\partial V}{\partial A}(r_{t+1}, w_{i,t+1}, A_{i,t+1}) . \tag{23.27}$$

in the neighborhood of $\mathrm{E}_t(r_{t+1}, w_{i,t+1}, A_{i,t+1})$, omitting the arguments in the various functions, yields

$$\begin{aligned} \lambda_{i,t+1} \simeq \frac{\partial V}{\partial A} &+ (r_{t+1} - \mathrm{E}_t r_{t+1}) \frac{\partial^2 V}{\partial A \partial r} + (w_{i,t+1} - \mathrm{E}_t w_{i,t+1}) \frac{\partial^2 V}{\partial A \partial w} \\ &+ (A_{i,t+1} - \mathrm{E}_t A_{i,t+1}) \frac{\partial^2 V}{\partial A^2} . \end{aligned} \tag{23.28}$$

Hence

$$\begin{aligned} \frac{\mathrm{Var}_t(\lambda_{i,t+1})}{2\mathrm{E}_t(\lambda_{i,t+1})^2} \simeq \theta_{Ar} \mathrm{Var}_t(r_{t+1}) &+ \theta_{Aw} \mathrm{Var}_t(w_{i,t+1}) \\ &+ \theta_{AA} \mathrm{Var}_t(A_{i,t+1}) + \text{covariance terms} , \end{aligned} \tag{23.29}$$

[8] On that account, it would be interesting to extend the model to allow for individual specific interest rates, and use information on household exposure to financial market risks for evaluating their labor supply behavior.

[9] Strictly speaking, $A_{i,t-1}$ should appear as an argument in function V. But examination of (23.28) shows that the corresponding terms are equal to zero, hence the simplification.

with

$$\theta_{Aj} \equiv \frac{1}{2}\left(\frac{\partial^2 V/\partial A\partial j}{\partial V/\partial A}\right)^2 \geq 0, \qquad j = r, w, A\ .$$

Equation (23.28) can also be used to calculate

$$\begin{aligned}\mathrm{Cov}_t\left(r_{t+1}, \lambda_{i,t+1}\right) = &\frac{\partial^2 V}{\partial A\partial r}\mathrm{V}_t\left(r_{t+1}\right) + \frac{\partial^2 V}{\partial A\partial w}\mathrm{Cov}_t\left(r_{t+1}, w_{i,t+1}\right)\\ &+\frac{\partial^2 V}{\partial A^2}\mathrm{Cov}_t\left(r_{t+1}, A_{i,t+1}\right)\ .\end{aligned}$$

For simplicity, we assume that $\mathrm{Cov}_t\left(r_{t+1}, \lambda_{i,t+1}\right)$ is constant in the sequel.

Using a first-order Taylor approximation to $\ln\lambda$ in the neighborhood of the realization $\left(r, w_i, A_i\right)_{t+1}$ gives

$$\begin{aligned}\mathrm{E}_t \ln\lambda_{i,t+1} &= \mathrm{E}_t \ln\frac{\partial V}{\partial A}\left(r_{t+1}, w_{i,t+1}, A_{i,t+1}\right)\\ &\simeq \ln\frac{\partial V}{\partial A} + \left(\mathrm{E}_t r_{t+1} - r_{t+1}\right)\frac{\partial^2 V/\partial A\partial r}{\partial V/\partial A}\\ &\quad + \left(\mathrm{E}_t w_{i,t+1} - w_{i,t+1}\right)\frac{\partial^2 V/\partial A\partial w}{\partial V/\partial A} + \left(\mathrm{E}_t A_{i,t+1} - A_{i,t+1}\right)\frac{\partial^2 V/\partial A^2}{\partial V/\partial A}\ .\end{aligned}$$

Thus,

$$\begin{aligned}\ln\lambda_{i,t+1} - \mathrm{E}_t \ln\lambda_{i,t+1} \simeq &-\eta_{Ar}\left(r_{t+1} - \mathrm{E}_t r_{t+1}\right) - \eta_{Aw}\left(w_{i,t+1} - \mathrm{E}_t w_{i,t+1}\right)\\ &-\eta_{AA}\left(A_{i,t+1} - \mathrm{E}_t A_{i,t+1}\right).\end{aligned} \tag{23.30}$$

where η_{AA} denotes the measure of absolute risk aversion in wealth (in terms of the indirect utility function), and

$$\eta_{Aj} \equiv -\frac{\partial^2 V/\partial A\partial j}{\partial V/\partial A}, \quad j = r, w, A\ ,$$

denote the change in marginal utility of wealth due to unanticipated changes in the explanatory variables. Notice that $\ln\lambda_{i,t+1} - \mathrm{E}_t \ln\lambda_{i,t+1}$ is uncorrelated with r_{t+1}, $w_{i,t+1}$ and $A_{i,t+1}$ under the assumption of rational expectations, see Hansen and Singleton (1982). In this case, former models that have neglected risk, and summed up $\ln\lambda_{i,t+1} - \mathrm{E}_t \ln\lambda_{i,t+1}$ with the residual term, mainly incur a loss in information and do not lead to an estimation bias.

Replacing (23.29) and (23.30) into (23.25) and adding a residual term u_{it} yields

$$\begin{aligned}\Delta\ln N_{it} = &\ \theta_0 + \eta\left(\rho - r_{t+1}\right) + \delta\Delta\ln w_{it}\\ &+\eta_r\left(r_{t+1} - \mathrm{E}_t r_{t+1}\right) + \theta_r \mathrm{Var}_t\left(r_{t+1}\right)\\ &+\eta_w\left(w_{i,t+1} - \mathrm{E}_t w_{i,t+1}\right) + \theta_w \mathrm{Var}_t\left(w_{i,t+1}\right)\\ &+\eta_A\left(A_{i,t+1} - \mathrm{E}_t A_{i,t+1}\right) + \theta_A \mathrm{Var}_t\left(A_{i,t+1}\right) + u_{it}.\end{aligned} \tag{23.31}$$

The parameter θ_0 comprises the covariance terms between the different types of risk. The parameters $\eta_r \equiv \eta\,(1-\eta_{Ar})$, $\eta_j \equiv -\eta\,\eta_{Aj}$, and $\theta_j \equiv -\eta\,\theta_{Aj}$ for $j = w, A$, reflect risk aversion with respect to variable j. It can directly be seen that risk has a negative impact on $\Delta \ln N_{it}$. The impacts of unanticipated changes in w, A are asymmetric. In the case where the marginal utility of wealth is decreasing in w_{it}, $\eta_{Aw} > 0$, and as $\eta > 0$, we have $\eta_w < 0$. Positive innovations in wages (*i.e.* $w_{i,t+1} > \mathrm{E}_t \widetilde{w}_{i,t+1}$) lead the individual to work less at $t+1$, whereas negative innovations have the opposite effect. In summary, unanticipated wage changes have the opposite impact to anticipated wage changes.

23.3.3 Empirical Results

With this framework it now becomes possible to investigate empirically the impacts of anticipated and unanticipated wage change on labor supply, and how individuals react to an increase in the variability of the lifetime wage profile. For instance, they could adopt a precautionary labor supply behavior in order to try to compensate the risk of a wage profile.

Pistaferri (2003) uses panel data from the Bank of Italy (Survey of Household Income and Wealth), which comprises subjective information for each individual on her anticipated wage profile and price inflation (implying cross-sectional variability in the real interest rates). The difference between observed and anticipated wage gives the unanticipated wage profile. Pistaferri's empirical specification is a special case of (23.31):

$$\Delta \ln N_{it} \approx \eta\,(\rho - \mathrm{E}_t r_{t+1}) + \delta \Delta \ln w_{it} + \beta\,\zeta_{it} + \gamma \mathrm{Var}_{t-1}(\zeta_{it}) + u_{it}\ ,$$

where

$$\zeta_{it} \equiv \ln w_{it} - \mathrm{E}_{t-1} \ln w_{it}\ .$$

Notice that in the neighborhood of zero, $\mathrm{Var}_{t-1}(\zeta_{it}) \approx \mathrm{Var}_{t-1}(w_{it}) / (\mathrm{E}_{t-1} w_{it})^2$.

Lich-Tyler (2002) sums up the unanticipated changes into a residual term v_{it}. Using the wealth identity (23.2), it can be seen that the wealth risk $\mathrm{Var}_{t-1}(A_{it})$ is driven by risk in the interest rate and risk in the future wage path. Assuming "that the wealth risk associated with a permanent wage change depends on the remaining work years of the individual and the amount of wage volatility" (Lich-Tyler, p.18), we write $\mathrm{Var}_{t-1}(A_{it}) = A_{i,t-1}^2 \mathrm{Var}_{t-1}(r_t) + \gamma_{wt} \mathrm{Var}_{t-1}(w_{it})\,(65-t)$. Putting things together, (23.31) boils down to

$$\begin{aligned}\Delta \ln N_{it} \;=\; & \theta_0 + \eta\,(\rho - r_t) + \delta \Delta \ln w_{it} + \gamma \mathrm{Var}_{t-1}(\zeta_{it}) \\ & + \alpha_r \mathrm{Var}_{t-1}(r_t) + \alpha_{rA} A_{i,t-1}^2 \mathrm{Var}_{t-1}(r_t) + \alpha_{wt} \mathrm{Var}_{t-1}(w_{it})\,(65-t) + v_{it}\ .\end{aligned}$$

Lich-Tyler relies on the PSID data for parameter estimation. In a first stage, the variance terms are estimated from the data, using various regressions.

Table 23.1 Labor supply estimates accounting for uncertainty and risk

	$\widehat{\eta}$	$\widehat{\delta}$	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{\alpha}_r$	$\widehat{\alpha}_{rA}$	$\widehat{\alpha}_{wt}$
Pistaferri (i)	0.59 (0.29)	0.70 (0.09)	−0.20 (0.09)	−0.11 (0.03)	–	–	–
Pistaferri (ii)	0.22 (0.18)	0.26 (0.05)	0.05 (0.06)	−0.05 (0.01)	–	–	–
Lich-Tyler	0.01 (0.04)	0.29 (0.09)	–	−0.13 (0.06)	−12.9 (4.6)	−0.05 (0.02)	−0.012 (0.005)

Some parameter estimates from both contributions are summarized in Table 23.1, estimated standard errors are given in parentheses.

The first line of Table 23.1, Pistaferri (i) shows the result of Pistaferri's basis estimates, whereas the second line, Pistaferri (ii), reports estimates of a model controlling for unemployment constraints. Further model estimations and robustness checks provide support for the first set of results. Pistaferri's estimate of the intertemporal elasticity of substitution is 0.70, which is somewhat higher than those usually reported. The last line of Table 23.1 gives the estimates obtained by Lich-Tyler. In this case, the elasticity of substitution of 0.29 is in line with those usually obtained from simpler models with the PSID data set. Pistaferri's estimate of the impact of wage innovation is significantly negative in his first model only. It implies that an unexpected 10% permanent upward shift in the wage profile decreases labor supply in all future periods by about 2.5%.

In all cases, the different types of risk have negative impact on the growth of labor supply. This finding is consistent with precautionary labor supply behavior. The estimates of γ are quite similar in the Pistaferri and Lich-Tyler studies. Whereas Pistaferri finds his estimate of wage risk to have a very limited impact on working behavior, Lich-Tyler's conclusions are quite different. His simulations show that wage risk can explain wide differences in working hour profiles (see his Fig. 23.2, p.35).

23.4 Voluntary and Involuntary Non-participation

Depending on the economic context, individuals are not always willing to work, or able to find a job, or able to work their desired amount of time. Taking this distinction into account is important for avoiding estimation biases. For instance, if after an increase in wages, a person loses her job, this does not mean that her labor supply decreases in wages.

23.4.1 Accounting for the Participation Decision

The prototype here is the paper by Heckman and MaCurdy (1980) which also presents the first estimation of a Tobit model on panel data.[10] The specification does not differ much from that of MaCurdy (1981) but now the individual considered is a married woman. Accounting for the participation decision is important because selecting only working individuals leads to a selection bias.

Separability between the leisures of husband and wife is assumed, and the specification chosen for the utility function is

$$U_{it}(C_{it}, L_{it}) = \gamma_{Cit} C_{it}^{\beta} + \gamma_{Lit} L_{it}^{\alpha_L} , \tag{23.32}$$

with $0 < \alpha_L < 1$, $0 < \beta < 1$. Maximization of (23.1) subject to (23.2), taking the possible nonparticipation into account, yields

$$\ln L_{it} = \begin{cases} \dfrac{1}{\alpha_L - 1}(\ln \lambda_{it} - \ln \alpha_L + \ln w_{it} - \ln \gamma_{Lit}) & \text{if} \quad L_{it} \leq \bar{L}, \\ \ln \bar{L} & \text{otherwise.} \end{cases} \tag{23.33}$$

The stochastic assumptions adopted are

$$\ln \gamma_{Lit} = Z_{it}\phi + \eta_{1i} + u_{1it}, \tag{23.34}$$

$$\ln w_{it} = X_{it}\psi + \eta_{2i} + u_{2it}, \tag{23.35}$$

$$\mathrm{E}u_{jit} = 0, \qquad Eu_{jit}u_{kis} = \delta_{ts}\sigma_{jk}, \quad j,\, k = 1,\, 2, \quad i = 1, ..., n, \quad s,\, t = 1, ..., T\ .$$

where η_{1i} and η_{2i} are individual fixed effects capturing unobserved heterogeneity in the specifications of $\ln \gamma_{Lit}$ and $\ln w_{it}$, and δ_{ts} is the Kronecker symbol. The error terms u_{1it} and u_{2it} are assumed independent of all other variables in the RHS of (23.34) and (23.35). The unobserved heterogeneity of the preference parameter γ_{Lit}, which reflects individuals' implicit valuation of leisure, may well be correlated with the unobserved heterogeneity η_{2i} driving the wage of individual i. In this case, wages are endogenous in (23.33). Substituting (23.34) and (23.35) into the labor supply function helps to circumvent this problem.

Heckman and MaCurdy consider the reduced form:

$$\ln L_{it} = \begin{cases} f_i + \dfrac{\rho - r}{\alpha_L - 1} t - Z_{it}\dfrac{\phi}{\alpha_L - 1} + X_{it}\dfrac{\psi}{\alpha_L - 1} + v_{it} & \text{if} \quad L_{it} \leq \bar{L}, \\ \ln \bar{L} & \text{otherwise.} \end{cases} \tag{23.36}$$

where

$$f_i = \frac{1}{\alpha_L - 1}(\ln \lambda_{i0} - \ln \alpha_L - \eta_{1i} + \eta_{2i})\ ,$$

and

[10] See also Heckman and MaCurdy (1982)

$$v_{it} = \frac{1}{\alpha_L - 1}(-u_{1it} + u_{2it}) \, .$$

Equations (23.35) and (23.36) are simultaneously estimated by ML, assuming normality for u_{1it} and u_{2it}.[11] Identification of all parameters requires exclusion restrictions between X and Z. The fixed effects are f_i in the hours equation and η_{2i} in the wage equation. The estimation can only be performed for women who worked at least once in the observed periods. Correction for the corresponding selection bias is found to have only a minor impact. Since asymptotic arguments are not justified in the time dimension (only eight waves), estimates of the fixed effects are not consistent and this leads in principle to the inconsistency of all the coefficients.[12] However, (i) Heckman (1981) performed Monte-Carlo simulations for fixed effects Probit with eight waves and found that the fixed effects Probit performed well when the explanatory variables were all strictly exogenous, (ii) Tobit should perform even better because it is a combination of Probit and linear regression. The fixed effects (incidental parameters) are estimated simultaneously with the parameters of interest through alternated iteration on both subsets of parameters.[13] Yet their economic interpretation is difficult because the influence of f is mixed with that of the time invariant variables in Z_t and the same holds for η_2 and the time invariant variables in X_t. Regressions of the fixed effects on those time invariant variables completes the picture and allows one to reach conclusions like the following: current-period household income (exclusive of the wife's earnings) has no significant impact on labor supply, in contrast to an 8 year average income (proxy for the permanent income).

Another study taking the participation decision into account is Jakubson (1988). The specification is the same as above but separate identification of ψ and ϕ is left aside and Jakubson specifies $X_t \equiv Z_t$. The model is thus considerably simplified and takes the Tobit form

$$\ln L_{it} = \begin{cases} f_i + \dfrac{\rho - r}{\alpha - 1} t + X_{it} \dfrac{\psi - \phi}{\alpha - 1} + v_{it} & \text{if} \quad L_{it} \leq \bar{L}, \\ \ln \bar{L} & \text{otherwise.} \end{cases} \tag{23.37}$$

Jakubson presents three approaches to the estimation of (23.37): simple pooling, treatment of f_i as a random effect taking into account the correlation with X (using Chamberlain's, 1984 approach) and, as before, treatment of f_i as a fixed effect. For the fixed effects, the considerations above still hold, while convergence for the random effects specification is ensured even for short panels as long as their stochastic specification is correct.

The main conclusions are: (i) the panel estimates (fixed or random effects) of the influence of children on labor supply are only about 60% of the cross section

[11] We do not mean to suggest that there are no alternatives to ML with joint normality in this context, and the interested reader is referred to Wooldridge (1995) and Dustmann and Rochina-Barrachina (2000) for some of these.

[12] That is, for $N \to \infty$.

[13] A computationally more efficient alternative is discussed by Greene (2004).

estimates, due to the neglect of individual effects in the latter; (ii) as concerns the life cycle hypothesis, like in the Heckman and MaCurdy study, current income does not have a significant influence in the fixed effects estimation, yet this does not hold true for random effects.

Disregarding the inconsistency problem associated with fixed effects here, and considering that sampling may be endogenous (one of the selection criteria being "stable marriage," see Lundberg, 1988) the fixed effects approach might seem preferable on *a priori* grounds. However, as we shall see in the following section, the entire specification is questionable.

Accounting for taxes is feasible in the framework discussed here, as documented by Laisney, Lechner, VanSoest and Wagenhals (1993). This study keeps the assumptions of explicit additivity of the intertemporal utility function and of intertemporal separability of the budget constraint. The specification postulates parallel within-period preferences, *i.e.*

$$U_{it}(C_{it}, L_{it}) = G_{it}[C_{it} + V_{it}(L_{it})] \,, \tag{23.38}$$

where G is an increasing function. This specification yields a useful benchmark, because the corresponding labor supply equation is independent of the marginal utility of wealth, λ_{i0} (and thus coincides with the Marshallian and the Hicksian labor supply equations). This clearly solves several of the econometric problems discussed above. Choosing a Box-Cox specification $V_{it}(L_{it}) = \gamma_{it}(L_{it}^{\alpha_L} - 1)/\alpha_L$ and keeping specifications (23.34) and (23.35) for the taste shifter γ_{it} and the gross wage w_{it} yields the labor supply equation

$$\ln L_{it} = \frac{1}{\alpha_L - 1}\left(\ln w_{it} + \ln[1 - \tau_t(w_{it} N_{it})] - \ln \gamma_{it}\right) + v_{it} \,, \tag{23.39}$$

where τ_t denotes the marginal tax rate, assumed here to vary only with earnings. This equation is very similar to (23.33), the specification of Heckman and MaCurdy (1980) apart from the fact that it does not include λ_{it} and ρ_i. However, as will be discussed in the next section, the Heckman–MaCurdy specification requires the restriction that G_{it} is the identity, so that, although the two labor supply equations are nested, the overall specifications are not. In the same spirit, it can be seen that the labor supply (23.61) and (23.33) of the Browning, Deaton and Irish (1985) and Heckman and MaCurdy (1980) specifications can be nested in the more general model

$$\frac{(L_{it})^{\kappa} - 1}{\kappa} = -\alpha_{it} - \delta \ln w^*(N_{it}) - \theta_1 \sqrt{\frac{1}{w^*(N_{it})}} - \delta \ln \lambda_{it} + v_{it} \,, \tag{23.40}$$

where $w^*(N_{it})$ denotes the real net (marginal) wage rate associated with N_{it}. The Browning et al. specification corresponds to the linear form $\kappa = 1$, whereas the Heckman–MaCurdy specification corresponds to the logarithmic specification obtained for the limiting case $\kappa = 0$, with $\theta_1 = 0$.

The model is estimated, taking the participation decision into account, using an unbalanced panel of married women drawn from the German Socio Economic Panel 1985–1989, using Mundlak's (1978) approach to modelling random effects for λ and Chamberlain's (1984) minimum distance estimator, whereby the first stage of the estimation procedure consists of (pseudo-) maximum likelihood simultaneous estimation of (23.34), (23.35) and (23.40). Following MaCurdy et al. (1990), the marginal tax rate is approximated by a smooth increasing function. A further distinctive feature of this study is that desired hours of work are used as the dependent variable, instead of effective hours of work. This weakens to some extent the critique of Tobit-type models of labor supply made by Mroz (1987).

23.4.2 Unemployment

Certainly one of the most questionable assumptions made so far is the assumption that unemployment is voluntary. Ham (1986) produces empirical evidence against this hypothesis in the context of life cycle models (see also Ashenfelter and Ham, 1979). Ham uses the following modification of MaCurdy's model. If an additional restriction consisting of a ceiling to the number of hours worked exists, and if T_u is the set of indices of the periods where this restriction holds for individual i, we have

$$\ln N_{it} < F_i + bt - \delta R_t + \delta \ln w_{it} + u_{it} \qquad \text{for} \quad t \in T_u \,, \tag{23.41}$$

$$\ln N_{it} = F_i + bt - \delta R_t + \delta \ln w_{it} + u_{it} \qquad \text{for} \quad t \notin T_u \,, \tag{23.42}$$

where F_i corresponds to a higher value of λ than when $T_u = \varnothing$: the profile of expected wages at each period is lower than in the absence of unemployment periods. Therefore, (23.13) will yield large residuals for $t \in T_u$ if unemployment is not the outcome of a free choice. The idea is then to estimate either

$$\ln N_{it} = F_i + bt - \delta R_t + \delta \ln w_{it} + \theta_1 U_{it} + u_{it} \tag{23.43}$$

or

$$\ln N_{it} = F_i + bt - \delta R_t + \delta \ln w_{it} + \theta_2 H^u_{it} + u_{it} \,, \tag{23.44}$$

where $U_{it} = 1$ if $t \in T_u$ and 0 otherwise, and H^u_{it} denotes yearly hours of unemployment. If the free choice assumption is correct, then θ_1 (or θ_2) will not significantly differ from zero. Otherwise one would expect negative values.

The free choice assumption is clearly rejected for both specifications (23.43) and (23.44), as well as for other specifications allowing for uncertainty, nonlinearity (with the additional term $(\ln w_{it})^2$), nonseparability, see (23.61), as well as for various assumptions on the covariance structure of the residuals. The results of these tests suggest modelling these restrictions explicitly. Lilja (1986) makes several proposals in this direction.

However, MaCurdy et al. (1990) criticizes Ham's argument and shows that θ_1 (or θ_2) significant in (23.43) or (23.44) is compatible with voluntary unemployment

caused by a lower wage offer w_{it} for $t \in T_u$: "The reasoning underlying the testing of exclusion restrictions in labour supply functions relies on the argument that wages fully capture the influences of demand-side factors in a supply decision. This reasoning is sound but the variable identified as relevant by intertemporal substitution theory is the offer wage; and the offer wage deviates from the observed market wage if unemployment occurs at all" (MaCurdy 1990, p. 228; see also Card, 1987, who interprets Ham's findings in favor of demand-side conditions as the main determinant of observed hours).

23.5 Alternative Parameterization and Implications

Browning (1986) and Blundell, Fry and Meghir (1990) point out that the specification of λ-constant systems, where λ, or $\ln\lambda$, appear additively and can be treated as an individual-specific effect turns out to be extremely restrictive in the models of MaCurdy (1981) and Browning et al. (1985). In this case, the labor supply functions share the form

$$g_i(N_{it}) = f_i(w_{it};\theta) + \delta \ln \lambda_{it} \tag{23.45}$$

where g_i and f_i are some functions, and θ and δ are parameters. After replacing $\ln\lambda_{it}$ by (23.8), first differentiation for individual i allows us to get rid of individual heterogeneity. The devastating consequence is that such intertemporal preferences are completely identified (up to a monotonic transformation) on a single cross section, given that some variation in the wages or prices can be observed. Thus, this type of specification hardly qualifies for exploiting panel data.

An alternative strategy consists in estimating the within-period preferences by eliminating λ, either directly between two goods or indirectly via the period budget equation, and then estimating the time preference rate ρ separately. The advantage is that no restriction on within-period preferences is required. Panel data are not absolutely necessary for this strategy: a time series of independent cross sections proves to be sufficient and even has some advantages in providing valid instrumental variables more easily, see Blundell, Fry and Meghir (1990). Blundell, Browning and Meghir (1994) give a good example of the application of this strategy to demands for goods. Four important panel studies on labor supply use this alternative strategy.

MaCurdy (1983) proposes to directly estimate the marginal rate of substitution functions. The first-order conditions (23.3) and (23.4) give

$$\frac{\partial U_{it}/\partial N_{it}}{\partial U_{it}/\partial C_{it}} = -w_{it}\ . \tag{23.46}$$

The advantage over estimating Marshallian demands is that this allows estimation of preferences that do not imply a closed-form expression for the demand functions. The estimation of (23.46) does not require a panel. A cross section with enough price variation, or indeed a time series of cross sections, can be sufficient.

In spite of this, MaCurdy chooses the restrictive form

$$U_{it} = G_{it}\left(U_{it}^*\right) = \xi_{it}\frac{(U_{it}^* + \nu)^{\sigma} - 1}{\sigma}\,, \tag{23.47}$$

with

$$U_{it}^* = \gamma_{it}\frac{(C_{it} + \theta_C)^{\alpha_C}}{\alpha_C} - \frac{(N_{it} + \theta_N)^{\alpha_N}}{\alpha_N}\,, \tag{23.48}$$

and

$$\xi_{it} = \exp[X_{it}\phi + \alpha_{it}]\,, \tag{23.49}$$

$$\gamma_{it} = \exp[X_{it}\psi + \varepsilon_{it}]\,. \tag{23.50}$$

The parameters $\phi, \psi, \sigma, \nu, \theta_C, \theta_N, \alpha_C$, and α_N are constant across individuals and over time. This utility function is still additive, yet no longer explicitly additive, and this form of U_t^* allows for several well-known special cases such as CES, addilog and Stone-Geary. The Frisch labor supply function corresponding to (23.47) is usually different from (23.45). There is no identification problem here since (23.49) and (23.50) are estimated in two different dimensions: (23.50) is estimated in the "individual" dimension and (23.49) in the "time" dimension. Equations (23.46) and (23.48) yield

$$\ln w_{it} = -X_{it}\psi + (\alpha_N - 1)\ln(N_{it} + \theta_N) - (\alpha_C - 1)\ln(C_{it} + \theta_C) - \varepsilon_{it}\,, \tag{23.51}$$

which provides consistent estimates (on a single cross section if desired) for ψ, α_N, α_C, θ_N and θ_C. Using those one can obtain γ_{it} by substitution of $X_{it}\psi + \varepsilon_{it}$ from (23.51) into (23.50). Estimates for the parameters σ and ϕ can be obtained as follows. Substitution of (23.3) into (23.6) gives

$$\ln\left(\frac{\partial U_{it}}{\partial C_{it}}\right) = r_{t+1} - \rho_i + \ln\left(\frac{\partial U_{i,t+1}}{\partial C_{i,t+1}}\right) + e_{i,t+1}\,. \tag{23.52}$$

The above specification leads to

$$\begin{aligned}\ln\frac{\partial U_{it}^*}{\partial C_{it}} - \ln\frac{\partial U_{i,t+1}^*}{\partial C_{i,t+1}} &= r_{t+1} - \rho_i - (X_{i,t+1} - X_{it})\phi \\ &\quad + (1-\sigma)\left[\ln(U_{it+1}^* + \nu) - \ln(U_{it}^* + \nu)\right] + e_{i,t+1}\,.\end{aligned} \tag{23.53}$$

Since estimates for U_{it}^* and $\partial U_{it}^*/\partial C_{it}$ are available from the parameter estimation of (23.51), specification (23.53) can be seen as a regression from which the still unknown parameters ϕ and σ of the monotonic transformation G_{it} can now be identified. Either time series or panel data contain all the information needed to estimate (23.53). Instrumental variables are necessary to take account of the endogeneity of U_{it}^* and $U_{i,t+1}^*$, and Pagan's (1984) method of correcting the variance of

the estimators would be advisable here, because estimated parameters are used in the construction of regressors as well as regressands in (23.53). Taking account of measurement errors in hours, wages or consumption would be difficult because such errors would contaminate ε_{it}, see (23.51), and would therefore produce nonlinear errors in the variables in (23.53).

The study of Blundell et al. (1993) of intertemporal labor supply of married women starts from the following Marshallian supply specification for within-period desired hours of work

$$N_{it} = \alpha(w_{it}, Z_{it}) - \beta(w_{it}, Z_{it})[S_{it} + a(w_{it}, Z_{it})] + u_{it} , \tag{23.54}$$

where w_{it} is the real marginal after tax wage rate, S_{it} is a measure for unearned income and Z_{it} is a vector of characteristics. This supply function can be derived by Roy's identity from the indirect utility function

$$V(w_{it}, S_{it}, Z_{it}) = \frac{1}{1+\rho(Z_{it})} \left[\left(\frac{S_{it} + a(w_{it}, Z_{it})}{b(w_{it}, Z_{it})} \right)^{1+\rho(Z_{it})} - 1 \right] , \tag{23.55}$$

with $\alpha(w_{it}, Z_{it}) = \partial a / \partial w_{it}$ and $\beta(w_{it}, Z_{it}) = (\partial b / \partial w_{it}) / b$. The parameters of $\rho(Z_{it})$ which do not show up in the labor supply function are estimated in a second stage, using GMM and relying on a procedure analogue to (23.53). Although the study discusses several different elasticities, we shall only comment on λ-constant elasticities of labor supply with respect to the net wage, computed at the means of various subsamples of employed women. These range between 0.57 for childless women with unemployed blue-collar husbands and 1.39 for women whose youngest child is at most two and whose husbands are employed white-collars, a subsample with typically low labor supply.

More on taxes: relaxing the intertemporal separability of the budget constraint. As pointed out by Blomquist (1985), capital taxation will usually break the intertemporal separability of the intertemporal budget constraint. When the constraints (23.2) are replaced by

$$A_{it} = (1 + r_t) A_{i,t-1} + w_{it} N_{it} - C_{it} - T(w_{it} N_{it} + r_t A_{i,t-1}) , \quad t = 1, \ldots, T ,$$

where the function T denotes the tax scheme. In this case, it is in general not possible to write the Frisch labor supply in function of an additive and constant λ-term which can be easily differentiated out. Ziliak and Kniesner (1999) consider instead a Marshallian labor supply function of the form

$$N_{it} = \alpha w_{it} + \delta A_{i,t-1} + \phi A_{it} + Z_{it}\gamma + \eta_i + \xi_{it} , \tag{23.56}$$

where η_i denotes an individual effect. The wealth measure used for A_{it} is a construct analogue to the familiar virtual income used in static models of labor supply with taxes. Notice that both current assets and assets in the previous period condition this labor supply function, which is thereby different from the pseudo supply function

$\widetilde{N}_{it}(A_{it} - (1+r_t)A_{i,t-1}, w_{it})$ of Sect. 23.2. Parameters reflecting the intertemporal utility function are then recovered in a second stage which is very similar to (23.52–23.53).

In contrast with Blundell et al., Ziliak and Kniesner use (balanced) panel data (PSID for 1978–1987), 532 continuously married, continuously working men aged 22–51 in 1978), in both stages of the estimation procedure. In the first stage, this has the advantage of allowing a better control of unobserved heterogeneity. Ziliak and Kniesner estimate (23.56) in first differences by optimal GMM assuming absence of autocorrelation in the process ξ_{it}, using internal instruments dated $t-2$ and other instruments dated $t-1$ and $t-2$. Estimation in the second stage is conducted on the same panel, with internal instruments dated $t-4$ and other instruments dated $t-3$ and $t-4$. A consistent estimator of the variance of the second stage estimator, taking into account the variability of the estimated quantities, is obtained following Newey (1984). Our impression is that potentially important efficiency gains might be obtained quite easily by (a) moving from a balanced to an unbalanced panel, which would substantially increase the number of observations, and (b) extending the instrument set by taking instruments dated *up to* the named dates rather than only instruments *at* the above dates.

Results from Step 1 show that the model conditioning on assets at two subsequent dates outperforms a model conditioning on savings. Two series of estimates are presented for Step 2, depending on whether the subjective discount rate, assumed constant over time, is allowed to vary over individuals or not. The former specification is the preferred one. Even though the λ-constant specification was not used for parameter estimation, the λ-constant elasticities of labor supply with respect to the gross wage are easily computed from (23.56). Mean λ-constant elasticities by wealth quartile vary between 0.14 for the lowest quartile and 0.20 for the highest. Recall that this represents the response to an expected wage change. By contrast the authors reckon that the average elasticity of labor supply with respect to an unexpected wage change will be roughly constant across wealth quartiles, at about 0.16**.** Ziliak and Kniesner also compute deadweight loss measures associated to four tax reforms, but reporting on these would take us too far off our track.

Errors in variables are thoroughly treated by Altonji (1986), using instrumental variables methods. Unfortunately, in order to obtain the required linearity, Altonji uses a version of MaCurdy's (1981) restrictive form, *i.e.* an explicitly additive within-period utility function

$$U_{it} = \frac{\gamma_{Cit}}{\alpha_C} C_{it}^{\alpha_C} - \frac{\gamma_{Nit}}{\alpha_N} N_{it}^{\alpha_N} , \tag{23.57}$$

where γ_{Cit} and γ_{Nit} are time-varying taste modifiers. The λ-constant demands are

$$\ln N_{it} = \mathrm{cst} + \delta_N [\ln w_{it} + \ln \lambda_{it} + t \ln(1+\rho) - \ln \gamma_{Nit}] , \tag{23.58}$$

$$\ln C_{it} = \mathrm{cst} + \delta_C [\ln \lambda_{it} + t \ln(1+\rho) - \ln \gamma_{Cit}] . \tag{23.59}$$

Rather than estimating (23.58) in first differences,[14] Altonji proposes substituting $\ln\lambda_{it} + t\ln(1+\rho)$ out of (23.58) and (23.59). He then assumes that the observations contain the measurement errors v^*_{Nit}, v^*_{Cit}, and e^*_{it}, and consist in $n^*_{it} = \ln N_{it} + v^*_{Nit}, c^*_{it} = \ln C_{it} + v^*_{Cit}$ and $w^*_{it} = \ln w_{it} + e^*_{it}$. Since w_{it} is not directly observed but is calculated by dividing period income by N_{it}, v^*_{Nit} is correlated with e^*_{it} but neither of the two will be correlated with v^*_{Cit}. Thus, we obtain the model:

$$n^*_{it} = \text{cst} + \delta_N w^*_{it} + \frac{\delta_N}{\delta_C} c^*_{it} + \delta_N \ln\frac{\gamma_{Cit}}{\gamma_{Nit}} + v^*_{Nit} - \delta_N e^*_{it} - \frac{\delta_N}{\delta_C} v^*_{Cit} \,. \tag{23.60}$$

The advantage over first differences is that the substitution using c^*_{it} does not bring lagged wages into the equation. Even more important perhaps, the assumption about expectations that was used above to motivate estimating first differences under uncertainty is now unnecessary. Instruments are used for w^*_{it} and c^*_{it}. The results do not differ much from MaCurdy's. See also Imai and Keane (2004) for a different treatment of the problem of errors in variables.

23.6 Relaxing Separability Assumptions

We now discuss studies relaxing within-period and between-period additive separability.

23.6.1 Relaxing Within-Period Additive Separability

When the within period utility function is additively separable, the Frisch demand functions satisfy the restrictions

$$\frac{\partial N}{\partial \lambda_{it}} \frac{\lambda_{it}}{N} = \frac{\partial N}{\partial w_{it}} \frac{w_{it}}{N} \quad \text{and} \quad \frac{\partial C}{\partial w_{it}} = 0 \,,$$

see (23.12). These restrictions are not simply a consequence of the functional form adopted in (23.12), indeed they characterize within-period additive separability. The importance of relaxing the assumption of separability between leisure and goods is indicated in Browning and Meghir (1991) who reject this assumption, testing it within a very general scheme using 1979–1984 FES data (time series of cross sections): preferences about goods are specified in a flexible way, with conditional cost functions where no behavioral assumption concerning labor supply or participation decision is needed.[15] Here we shall be concerned only with relaxing the assumption of additive separability between the two "goods" leisure and aggregate consumption.

[14] Yet this is done for comparison.

[15] Yet their model is not cast in the life cycle framework and the implications of their study for life cycle models should be elucidated.

Browning et al. (1985) estimate the following specification in first differences:

$$N_{it} = \alpha_1(a_{it}) + \delta \ln w_{it} + \theta_1 \sqrt{\frac{1}{w_{it}}} + \delta \ln \lambda_{it} \,, \tag{23.61}$$

$$C_{it} = \alpha_2(a_{it}) - \theta_2 \sqrt{w_{it}} + \beta \ln \lambda_{it} \,, \tag{23.62}$$

where a_{it} is a vector of household characteristics. Symmetry of the demand functions implies that $\theta_1 = \theta_2 = \theta$ and within-period additive separability is equivalent to $\theta = 0$. Browning et al. (1985) estimate the equations separately, *i.e.* they do not enforce the identity $\theta_1 = \theta_2$, as would be feasible in this context since there is no adding-up restriction (in contrast with a Marshallian demand system). However, they find θ_1 and θ_2 to be significantly different from zero and to have opposite signs, which makes the entire specification appear questionable. Note that, although Browning et al. consider aggregate consumption, no problem arises from working with several consumption goods. Yet, durables should be given special attention, as they might be more properly treated as assets.

So far we have focused on the preferences of an individual. In practice, however, economists often work with "household preferences". One of the many reasons for doing this is the difficulty of isolating individual from household consumption in survey data. Another assumption, which is necessary for the validity of the specifications that we have considered so far is the separability of the labor supplies of the different potential earners in a household. If it holds, the earnings of the other household members can be accounted for in A_{it}, because then the influence of hours and wages of other household members boils down to a pure income effect. Otherwise the model is misspecified.

23.6.2 Relaxing Intertemporal Separability in Preferences

Although relaxing this assumption is no easy task, it is important because all the studies that test the assumption clearly reject it. If the estimation results are to be used in policy analysis, the specification must produce interpretable parameters and not merely a separability test. In this respect, it seems difficult to simultaneously model the multiple reasons that lead to the rejection of separability. Most empirical studies therefore concentrate on only one of these aspects. The modelling of partial adjustment, rational habit formation and human capital accumulation in an optimization scheme over the life cycle is such a feasible extension.

Yet, before turning to structural models relaxing the intertemporal separability assumption, it is interesting to discuss the results of a VAR approach to modelling the relationship between wages and hours of work using panel data. As a prototype for this kind of approach we will focus on the study by Holtz-Eakin, Newey and Rosen (1988), but also refer the reader to Abowd and Card (1989).

Holtz-Eakin et al. analyze a sample of 898 males from the Panel Study of Income Dynamics (PSID) over 16 years. They estimate linear equations for wages and

hours, with lags of equal lengths on both wages and hours on the right hand side of each equation, and individual effects. Note that the equation on hours does not nest the simple life-cycle model of MaCurdy (1981) since the contemporaneous wage is excluded and no serial correlation is allowed. By contrast, the form of the wage equation could be justified by human capital considerations. However, attempts at interpreting these reduced form equations are not in line with the VAR approach. The model of Holtz-Eakin et al. does not a priori impose the stationarity of the coefficients over time, not even for the individual effect. The estimation strategy relies on GMM, combined with quasi-differencing along the lines of Chamberlain (1984, p. 1263) in order to eliminate the individual effect while allowing for nonstationarity. Errors in variables are easily dealt with in this linear GMM framework, but again under the restrictive assumption of no serial correlation. Starting with a maximum lag length of three periods (involving four lags of the original variables in the quasi-differenced equations) parameter stability is rejected for none of the two equations, and the analysis proceeds more simply with first differences. The next step concerns testing the lag-length, and the assumption that one lag is sufficient to describe the data is rejected in no equation at the 1% level, but rejected in the hours equation at the 5% level.

Furthermore, one cannot reject the assumption that lagged hours could be excluded from the wage equation. The same holds for lagged wages in the hours equation, when using only one lag, but not if two lags are retained (an argument in favor of nesting the noncausality test within the hypothesis about the lag length is that in this way the test statistics turn out to be asymptotically independent, which facilitates pin-pointing the reasons for rejection of the joint hypothesis). Tests for measurement error bias are constructed using internal instruments in the simple first-order autoregressive models, in order to increase the power of the test. The assumption of absence of measurement error cannot be rejected at the 5% level, but there is evidence that the test may have low power in this instance. Most results are qualitatively, and, what is more surprising, quantitatively replicated on a sample from the National Longitudinal Survey (NLS). The authors conclude (p. 1393): "Our empirical results are consistent with the absence of lagged hours in the wage forecasting equation, and thus with the absence of certain human capital or dynamic incentive effects. Our results also show that lagged hours are important in the hours equation, which is *consistent with the alternatives to the simple labour supply model that allow for costly hours adjustment or preferences that are not time separable* [our emphasis]. As usual, of course, these results might be due to serial correlation in the error term or functional form misspecification".

23.6.2.1 Rational Habit Formation

Bover (1991) estimates a rational habits model in a certainty framework with a minimum amount of replanning. The salient feature of her approach is that the model specification is constructed in such a way that it allows for an explicit expression of the marginal utility of wealth λ, as a function of future wages, initial wealth,

the (constant) interest rate, and preference parameters. The advantage of such an expression is that it allows a direct analysis of wealth effects on intertemporal labor supply (see Card, 1994, for the potential importance of such effects), whereas the approach of MaCurdy (1981) allows such an analysis only in a very indirect and unsatisfactory way. However, this comes at a large cost, as we shall see. In period t the individual maximizes

$$\sum_{t=1}^{T} \frac{1}{(1+\rho)^{t-1}} \left[(1-\beta_{it}) \ln(C_{it} - \gamma_c) + \beta_{it} \ln(\gamma_N + \phi N_{i,t-1} - N_{it})\right] \tag{23.63}$$

subject to (23.2). The parameter ϕ now measures the habit persistence. The Stone-Geary specification (23.63) was also used by Ashenfelter and Ham 1979 in order to derive an explicit expression for λ_{it} under perfect foresight. The novel feature here lies in the relaxation of the intertemporal separability assumption through the rational habit formation assumption. (In a previous paper Bover, 1986, considered two alternative models, one with partial adjustment and one with myopic habit formation, which did not take account of all direct and indirect influences of current labor supply on future decisions, as the rational habit formation model does, but she found all these models to be empirically indistinguishable.)

Defining $N_{it}^* = N_{it} - \phi N_{i,t-1}$ and $w_{it}^* = \sum_{j=0}^{T-t}(1+r)^{-j}\phi^j w_{i,t+j}$ allows one to rewrite (23.63) and (23.2) in the usual form of a separable intertemporal utility function with arguments $\{N_{it}^*, C_{it}\}_{t=1,\ldots,T}$ and an additively separable intertemporal budget constraint. The corresponding Frisch demands are linear in λ_{it} and the expression of the latter is obtained by substituting these into the budget constraint. The reason for the subscript t in λ_{it} is the replanning that takes place at each period, when the individual forms new predictions about his wage profile. The somewhat arbitrary assumption here is that each individual's future wages lie on a specific linear time trend, and that the individual learns more about the two coefficients of this relationship as more time passes by. This is disturbing, because if the relationship were deterministic, two observations would suffice to pin it down without any error, and if not we have uncertainty about future wages, whereas the derivation of λ_{it} assumed that w_{it}^* is known.

This specification yields a nonlinear model where the dynamics are only present in the error term. The model can be exactly linearized through transformations of the exogenous variables on the one hand, and the parameters on the other. The error specification is of the error components type with the unobserved heterogeneity subsumed in a time-invariant individual effect. Bover estimates the dummy variable model with unrestricted covariance for the residual error term, including also time dummies and using instruments to cope with potential endogeneity and measurement error problems concerning the wage variable. The instruments used have the property that they are strictly exogenous conditional on the individual effect. A χ^2 test of the overidentifying restrictions leads to no clear-cut rejection of the specification. The results show that lagged hours have a significant effect on the current decision.

While Bover substitutes the marginal utility of wealth in the Euler equation with a very special assumption about the wage path, Hotz, Kydland and Sedlacek (1988) (HKS) consider the stochastic Euler equations, characterizing the first-order conditions of the dynamic optimization problem. This strategy allows to consider more general specifications for the utility functions. In period t the individual maximizes

$$\mathrm{E}_t \sum_{t=1}^{T} \frac{1}{(1+\rho)^{t-1}} U_{it}(C_{it}, L_{it} + \alpha a_{it}) \,, \tag{23.64}$$

with

$$a_{it} = (1-\eta) a_{i,t-1} + L_{i,t-1} \,,$$

subject to (23.2). As before L_{it} denotes leisure. This specification nests intertemporal separability $(\alpha = 0)$ and the models of Johnson and Pencavel (1984) and Bover (1986, 1991), where only the labor supply of the previous period does play a role in the preferences of the current period $(\eta = 1)$.

In order to avoid misspecification, stemming from potential endogeneity of wages, HKS only use the Euler equation for consumption. They specify U_{it} to be translog and separately estimate the parameters for two age groups. Since parameters α and η are identified under the maintained assumption of no contemporaneous additive separability between $L_{it} + \alpha a_{it}$ and C_{it}, this allows testing the form of the intertemporal nonseparability in preferences. Moreover, a score test of the wage exogeneity is offered. HKS also explain how to cope with a certain degree of correlation between individuals through macroeconomic shocks or regional variables. Using a sample of 482 men from the PSID, they reach the following conclusions. The estimated parameters α and $(1-\eta)$ are positive and well determined and therefore intertemporal separability is rejected, and not only $L_{i,t-1}$ but also leisure decisions in previous years have a direct influence on current decisions. The (within period) separability between $L_{it} + \alpha a_{it}$ and C_{it} in the translog utility function is also rejected, as is exogeneity of the wages. A slightly disturbing result is the negativity of the estimated rate of time preference.

The theoretical setting (Euler equation) implies orthogonality between the residual at time t and all the information available up to $t-1$. Thus, in GMM estimation, all variables dated $t-1$ or earlier qualify in principle as instruments for the equation dated t. This implication of theory can be tested by a χ^2–test of overidentifying restrictions using two sets of instruments, where one is restricted to strictly exogenous instruments. HKS conduct such a test and do not reject the null of orthogonality.

23.6.2.2 Human Capital Formation

Hotz et al. and Bover assume that the wage path is not influenced by the hours decision, thus assuming intertemporal separability in the budget constraint. By contrast, Shaw (1989) and Imai and Keane (2004) relax that assumption, *i.e.* they allow for nonseparability in the budget constraint (but not in the preferences). For Shaw, $\alpha = 0$ in (23.64). The budget restriction is as before given by (23.2). However, Shaw

defines the real wage w_{it} as the product $R_{it}K_{it}$ of the human capital stock K_{it} and its rental rate R_t and chooses a quadratic approximation f for the relationship between $K_{i,t+1}$ on the one side and K_{it} and N_{it} on the other side, which yields the atypical earnings function

$$\frac{w_{i,t+1}}{R_{i,t+1}} = f\left(N_{it}, \frac{w_{it}}{R_{it}}\right) . \tag{23.65}$$

Although Shaw considers a timewise separable *direct* utility function, this last equation, which makes future wages an increasing function of the current wage, renders the *indirect* utility function nonseparable. The first-order optimality condition with respect to leisure now reads:

$$\frac{\partial U_{it}}{\partial L_{it}} = \lambda_{it} w_{it} \left(1 + R_{it} \frac{\partial f^{-1}}{\partial N_{it}} \frac{N_{it}}{w_{it}}\right), \quad t = 1, \ldots, T , \tag{23.66}$$

which differs from the timewise separable optimality condition (23.4). When working today increases future wages, this leads individuals to work more as predicted by timewise separable models.[16] Imai and Keane (2004) provide a further contribution along these lines.

Shaw specifies U_{it} to be translog (as in HKS). Preference parameters are estimated by GMM using the orthogonality conditions in the stochastic Euler equations. This contrasts with Imai and Keane (2004) who solve the stochastic dynamic programming problem backwards.

Shaw's conclusions are as follows. The rental rate of human capital varies considerably over time and the number of hours worked has a strong influence on future wages. This result offers a possible explanation for the misspecification of the usual static earnings function. Because of the model structure and especially the fact that the nonlinearity is within the budget constraint, the overall implications of the model can only be evaluated by simulation. This reveals that the intertemporal elasticity of labor supply is not constant as is usually assumed in static models, but instead rises over the life cycle. Her model is estimated over a samples of 526 men from the PSID. Due to the high degree of nonlinearity in the Euler equations, Shaw does not handles measurement errors or unobserved heterogeneity. In particular, the presence of unobserved heterogeneity is problematic as it can bias the conclusions about state dependence in dynamic models (see Chamberlain, 1984).

A reason why the models of Shaw and HKS have been estimated with male rather than with female labor supply may be that the estimation method used does not readily extend to discrete data. Altug and Miller (1991) propose a solution to that problem. We shall not go into the details of their paper, but it seems worth mentioning that this is a very sophisticated and innovative study, which also

[16] Notice that for the alternative specification allowing the earning function f to depend on the *cumulative* hours of work (and concave in this variable), there are especially young individuals who have incentives to work more whereas older individuals for which human capital investment become less attractive – given the fact that retirement is nearer than for young individuals – to work less.

considerably improves upon the treatment of aggregate shocks adopted by the two studies just mentioned. The main drawback is that estimation of the model of Altug and Miller is intricate, combining GMM with simulation of participation probabilities and iterative estimation of Euler equations, including nonparametric regressions at each iteration. In short, it required the use of a supercomputer. Another drawback, a theoretical one, is that the model heavily relies on the assumption that actual hours of work differ from expected or contracted hours of work in a stochastic manner. While this may be attractive for some occupations (think of academics), it is much less convincing for most others. To our knowledge, this is the only study of female labor supply allowing for nonseparability both in the preferences and in the budget constraint.

The study of Eckstein and Wolpin (1989), which is based on explicit solution of the dynamic optimization problem facing individuals rather than on the exploitation of first-order conditions, shares this generality but restricts attention to the participation decision and disregards aggregate shocks. Hence it does not exactly fit the framework of this survey.[17] It has however inspired the work of Imai and Keane (2004), to which we now turn. The main goal of their study is to reconcile the microeconometric evidence on the *ies* with the higher values adopted by macroeconomists in the calibration of real business cycle models. Their framework assumes both intertemporal and within period additive separability of preferences, but it allows for on-the-job human capital accumulation to affect the wage path, which breaks intertemporal separability in the budget constraint, as in Shaw (1989). It also allows for measurement errors in wages, labor supply, and assets in a maximum likelihood framework with fully parametric distributional assumptions. Missing data on assets are also handled through both distributional assumptions and the intertemporal budget constraint. The functional forms adopted for the subutility functions from consumption and leisure are the same as in MaCurdy (1981), except for the presence of age effects in the former. The intertemporal budget constraint is again (23.2) where t denotes the age of the individual, and the real wage rate w_t, assimilated with the human capital stock, evolves according to

$$w_{t+1} = g\left(N_t, w_t, t\right) \varepsilon_{t+1} \ ,$$

where ε_{t+1} is a wage shock and g is a deterministic function of hours worked and human capital at age t, and age itself.

Imai and Keane argue that neglecting human capital accumulation biases *ies* estimates towards zero. On the one hand, as the wage increases over the life-cycle, the substitution effect leads to an increase in labor supply. On the other hand, concavity of the value function in human capital lowers the rate of return to human capital investment and reduces the incentive to supply labor. The combination of the two effects leads to a fairly flat hours-wage profile, and attributing this to the substitution effect only leads to an underestimation of the *ies*.

Indeed, estimating their model on a fairly homogeneous sample of 1000 randomly chosen white males from the 1979 cohort of NLSY observed at ages 20–36

[17] A summary of that study is in the 1996 version of this survey.

and continuously reporting positive yearly hours of work, Imai and Keane obtain an estimate of the *ies* of 3.82 with a very small standard error.[18] They then simulate data from their model over the life-cycle up to age 65 and estimate the *ies* on various subsets using the OLS and IV methods of MaCurdy and Altonji. The results show that the estimated *ies* is much lower in these estimates than the true *ies* in the simulated data, and that estimates are particularly low for individuals in the 20–36 age group, underscoring the fact that the human capital component of the return to labor supply is much greater for the young. Indeed, IV results obtained from the original data yield an *ies* below 0.3, more than ten times smaller than the ML estimate.

Imai and Keane (2004) estimate the marginal rate of substitution between consumption and labor supply, which corresponds to $w_{it}\ [1+R_{it}\left(\partial f^{-1}/\partial N_{it}\right)(N_{it}/w_{it})]$ in (23.66). Their results range from about $2w_{it}$ for 20 years old individuals to w_{it} for 60 years old. This means that the effective wage is higher than market wages w_{it} due to high return in human capital, which induces young people to work more (at given wage) than predicted by (23.4).

23.7 Conclusion

Taking stock, we can draw the following conclusions. Firstly, in our opinion, there has so far still being too little emphasis on the relaxation of ad hoc functional form assumptions. In a way, this is understandable, because researchers have been busy introducing and manipulating new and sometimes complex econometric methods. Yet it is disturbing to see how popular the additively separable Box-Cox type specification has remained over the 25 past years, even in studies allowing much more flexible approaches. The greater flexibility of the alternative to Frisch demands, consisting in separate estimation of within-period and intertemporal preference parameters, has not yet been used fully in life-cycle labor supply studies. Secondly, given the small sample sizes and the more or less pronounced arbitrariness of the selection, most of the studies we have discussed definitely have a methodological rather than a substantive character.

Before closing this chapter we would still like to point out a series of papers which do not completely fit under its heading but contribute to the understanding of labor supply reactions. Blundell, Duncan and Meghir (1998) use past fiscal reforms in order to estimate labor supply responses. For estimation they rely on a series of cross-sections but their innovative approach can easily be adapted to panel data. A growing body of literature relies on daily information on wages and working time for particular worker groups to investigate the sensitivity of working time to wages: cabdrivers have been considered by Camerer, Babcok, Loewenstein and Thaler (1997) and by Farber (2005), stadium vendors by Oettinger (1999), bicycle messengers by Fehr and Götte (2007). This type of data exhibits two important

[18] The estimation method is too complex to be described in any detail in a survey. It entails several clever approximations aiming at reducing the number of evaluation points and the dimension of the optimisation space.

advantages over usual panel data: these workers choose daily the number of working hours they want to work, and daily variations of their hourly wage can reasonably be considered as transitory changes. Their results tend to suggest a negative relationship between wages and working hours. Finally, there is a burgeoning literature on the estimation of *collective* models of household labor supply, *i.e.* models where the existence of autonomous decision makers within the household is explicitly acknowledged, and the central assumption is that household allocations are Pareto efficient. For instance, Blundell, Chiappori, Magnac and Meghir (2005) estimate a static model on repeated cross-sections, focusing on the participation/nonparticipation decision of the husband while allowing free choice of hours for the wife (including nonparticipation). The longitudinal information contained in panel data allows the study of intertemporal household allocations in the collective framework, as exemplified by Mazzocco (2007).

Acknowledgments This is an update on Chap. 28 of the 1996 edition of this book. Besides incorporating new references, we also modified the focus. The reader is referred to the former version for a summary of quantitative results, as well as for data issues. We would like to thank Richard Blundell, Tom Kniesner, Gauthier Lanot, Michael Lechner, Friedhelm Pfeiffer, Jörn-Steffen Pischke, Jean-Marc Robin and Gerhard Wagenhals for comments.

References

Abowd, J. and D. Card (1989): "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57(2), 411–445.

Altonji, J.G. (1986): "Intertemporal Substitution in Labour Supply: Evidence from Micro-Data," *Journal of Political Economy*, 94 (3.2), S176–S215.

Altug, S. and R.A. Miller (1990): "Household Choices in Equilibrium," *Econometrica*, 58, 543–570.

Altug, S. and R.A. Miller (1991): "Human Capital Accumulation, Aggregate Shocks and Panel Data Estimation," Discussion Paper 9128, CentER, Tilburg University.

Ashenfelter, O. and J. Ham (1979): "Education, Unemployment and Earnings," *Journal of Political Economy*, 87, S99–S116.

Blomquist, S. (1985): "Labour Supply in a Two-Period Model: The Effect of a Nonlinear Progressive Income Tax," *Review of Economic Studies*, 52, 514–524.

Blundell, R.W. (1987): "Econometric Approaches to the Specification of Life-Cycle Labor Supply and Commodity Demand Behaviour," *Econometric Reviews*, 6(1), 147–151.

Blundell, R.W. (1988): "Consumer Behaviour: Theory and Empirical Evidence - A Survey," *Economic Journa*l, 98, 16–65.

Blundell, R.W and T. MaCurdy (1999): "Labor Supply: A Review of alternative Approaches," in *Handbook of Labor Economics*, Volume 3A, O. Ashenfelter and D. Card (eds.), Elsevier Science, Amsterdam.

Blundell, R.W. and C. Meghir (1990): "Panel Data and Life Cycle Models," Chapter 2 in *Panel Data and Labour Market Studies*, J. Hartog, G. Ridder and J. Theeuwes (eds.), North Holland, Amsterdam.

Blundell, R.W., M. Browning and C. Meghir (1994): "Consumer Demand and the Life-Cycle Allocation of Household Expenditures," *Review of Economic Studies***,** 61, 57–80.

Blundell, R.W, A. Duncan and C. Meghir (1998): "Estimating Labor Supply Responses Using Tax Reforms," *Econometrica*, 66, 827–861.

Blundell, R.W., V. Fry and C. Meghir (1990): "Preference Restrictions in Microeconometric Models of Life Cycle Behaviour," Chapter 2 in *Microeconometrics: Surveys and Applications*, J.P. Florens, M. Ivaldi, J.J. Laffont and F. Laisney (eds.), Basil Blackwell, Oxford.

Blundell, R.W, C. Meghir and P. Neves (1993): "Labour supply and intertemporal substitution," *Journal of Econometrics*, 59, 137–160.

Blundell, R.W, P.-A. Chiappori, T. Magnac and C. Meghir (2005): "Collective Labour Supply: Heterogeneity and Nonparticipation," *IZA Discussion Paper* 1785.

Bover, O. (1986): "Some Dynamic Life Cycle Models of Labour Supply Estimated From Panel Data," Oxford Applied Economics Discussion Papers, 13.

Bover, O. (1989): "Estimating Intertemporal Labour Supply Elasticities Using Structural Models," *Economic Journal*, 99, 1026–1039.

Bover, O. (1991): "Relaxing Intertemporal Seperability: A Rational Habits Moel of Labour Supply Estimated from Panel Data," *Journal of Labour Economics*, 9, 85–100.

Browning, M. (1986): "The costs of using Frisch demand functions that are additive in the marginal utility of expenditure," *Economics Letters*, 21(3), 205–207.

Browning, M. and C. Meghir, (1991): "The Effects of Labour Supply on Commodity Demands," *Econometrica*, 59, 925–952.

Browning, M., A. Deaton and M. Irish (1985): "A Profitable Approach to Econometric Approaches to Labor Supply and Commodity Demands over the Life Cycle," *Econometrica*, 53, 503–543.

Camerer, C., L. Babcok, G. Loewenstein and R. Thaler, 1997, "Labor Supply of New York City Cabdrivers: One Day at a Time," *Quaterly Journal of Economics*, 112, 407–441.

Card, D. (1987): "Supply and Demand in the Labor Market," Princeton University Industrial Relations Section, Working Paper No. 228.

Card, D. (1994): "Intertemporal Labor Supply: an Assessment", in *Advances in Econometrics, Sixth World Congress*, C. Sims (ed.), Cambridge University Press, New York.

Chamberlain, G. (1984): "Panel Data," Chapter 22 in *Handbook of Econometrics*, Vol II, Z. Griliches and M.D. Intriligator (eds.), North-Holland, Amsterdam.

Conway, K. S. and T. J. Kniesner (1994): "Estimating labor supply with panel data," *Economics letters*, 44, 27–33.

Dustmann C. and M. E. Rochina-Barrachina (2000): "Selection correction in panel data models: an application to labour supply and wages," *IZA Discussion Paper* 162.

Eckstein, Z. and K.I. Wolpin (1989): "Dynamic Labour Force Participation of Married Women and Endogenous Work Experience," *Review of Economic Studies*, 56, 375–390.

Farber, Henry S. (2005): "Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers," *Journal of Political Economy*, 119, 46–82.

Fehr, E. and L. Götte (2007): "Do Workers Work More When Wages are High", *American Economic Review*, 97(1), 298–317.

Greene, W.H. (2004): "Fixed Effects and the Incidental Parameters Problem in the Tobit Model," *Econometric Reviews*, 23(2), 125–148.

Ham, J.C. (1986): "Testing Whether Unemployment Represents Intertemporal Labour Supply Behaviour," *Review of Economic Studies*, 53(4), 559–578.

Hansen, L. P. and J. S. Singleton (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 50, 1269–1286.

Heckman, J.J. (1981): "The Incidental Parameters Problem and the Problem of initial Conditions in Estimating a Discrete Time - Discrete Data Stochastic Process and Some Monte-Carlo Evidence," in *Structural Analysis of Discrete Data*, C. Manski and D. McFadden, (eds.), MIT Press, Cambridge, MA, 179–195.

Heckman, J.J. and T.E. MaCurdy (1980): "A Life Cycle Model of Female Labour Supply," *Review of Economic Studies*, 47, 47–74.

Heckman, J.J. and T.E. MaCurdy (1982): "Corrigendum on A Life Cycle Model of Female Labour Supply," *Review of Economic Studies*, 49, 659–660.

Heckman, J.J. and T.E. MaCurdy (1986): "Labor Econometrics," Chapter 32 in *Handbook of Econometrics*, Vol III, Z. Griliches and M.D. Intriligator (eds.), North-Holland, Amsterdam.

Henley, A. (2004): "House Price Shocks, Windfall Gains and Hours of Work: British Evidence", *Oxford Bulletin of Economics and Statistics*, 66(4), 439–456.

Holtz-Eakin, D., W. Newey and H.S. Rosen (1988): "Estimating Vector Autoregressions with Panel Data," *Econometrica*, 56, 1371–1395.

Hotz, V.J., F.E. Kydland and G.L. Sedlacek (1988): "Intertemporal Preferences and Labour Supply," *Econometrica*, 335–360.

Imai, S. and M. P. Keane (2004): "Intertemporal Labor Supply and Human Capital Accumulation.," *International Economic Review*, 45(2), 601–641.

Jakubson, G. (1988): "The Sensitivity of Labor Supply Parameter Estimates to Unobserved Individual Effects: Fixed- and Random-Effects Estimates in a Nonlinear Model Using Panel Data," *Journal of Labor Economics*, 6(3), 302–329.

Johnson, T.R. and J.H. Pencavel (1984): "Dynamic Hours of Work Functions for Husbands, Wives and Single Females," *Econometrica*, 52, 363–389.

Joulfaian, D. and M. O. Wilhelm (1994): "Inheritance and Labor Supply," *Journal of Human Resources*, 29, 1205–1234.

Killingsworth, M.R. (1983): *Labor Supply*, Cambridge University Press, Cambridge.

Killingsworth, M.R. and J.J. Heckman (1986): "Female Labor Supply: A Survey," Chhapter 2 in *Handbook of Labor Economics*, Vol I, O. Ashenfelter and R. Layard, (eds.), North-Holland, Amsterdam.

Laisney, F., M. Lechner, A. van Soest and G. Wagenhals (1993): "A Life Cycle Labour Supply Model with Taxes Estimated on German Panel Data : The Case of Parallel Preferences," *Economic and Social Review*, 24, 335–368.

Lich-Tyler, S. (2002): "Life-cycle labor supply under uncertainty," presented at the 2002 North American Econometric Society Winter Meeting.

Lilja, R. (1986): *Econometric Analyses of Family Labour Supply over the Life Cycle Using US Panel Data*, The Helsinki School of Economics, Helsinki.

Lundberg, S.J. (1988): "Labor Supply of Husbands and Wives: A Simultaneous Equations Approach," *Review of Economics and Statistics*, 70(2), 224–235.

MaCurdy, T.E. (1981): "An Empirical Model of Labor Supply in a Life-Cycle Setting," *Journal of Political Economy*, 89, 1059–1085.

MaCurdy, T.E. (1983): "A Simple Scheme for Estimating an Intertemporal Model of Labor Supply and Consumption in the Presence of Taxes and Uncertainty," *International Economic Review*, 24, 265–290.

MaCurdy, T.E. (1990): "Appraising Tests of the Intertemporal Substitution Hypothesis," in *Panel Data and Labour Market Studies*, J. Hartog, G. Ridder and J. Theeuwes (eds.), North Holland, Amsterdam.

MaCurdy, T.E., D. Green and H. Paarsch (1990): "Assessing Empirical Approaches for Analyzing Taxes and Labor Supply," *Journal of Human Resources*, 25, 413–490.

Mazzocco, M. (2007): "Household Intertemporal Behavior: a Collective Characterization and a Test of Commitment", *Review of Economic Studies*, 74(3), 857–895.

Mincer, J. (1962): "Labor Force Participation of Married Women: A Study of Labor Supply," *Aspects of Labor Economics*, NBER, Princeton University Press, Princeton, N.J., 63–97.

Mroz, T.A. (1987): "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765–799.

Mundlak, Y. (1978): "On the Pooling of Time-Series and Cross-Section Data," *Econometrica*, 46, 69–85.

Newey, W.K., (1984): "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters,* 14, 201–206.

Oettinger G. S. (1999): "An Empirical Analysis of the Daily Labor Supply of Stadium Vendors," *Journal of Political Economy*, 107, 360–392.

Pagan, A. (1984): "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, 25, 221–247.

Pencavel, J. (1986): "Labor Supply of Men: A Survey," in O. Ashenfelter and R. Layard (eds.), *Handbook of Labor Economics*, North-Holland, Amsterdam.

Pistaferri, L. (2003): "Anticipated and unanticipated wage changes, wage risk and life cycle labor supply," *Journal of Labor Economics*, 21, 729–754.

Shaw, K. (1989): "Life Cycle Labor Supply with Human Capital Accumulation," *International Economic Review*, 30(2), 431–457.

Wooldridge, J. M. (1995), "Selection corrections for panel data models under conditional mean independence assumptions,"emph Journal of Econometrics, 68, 115–132.

Ziliak, J.P. and T.J. Kniesner (1999): "Estimating Life-Cycle Labor Supply Tax Effects," *Journal of Political Economy*, 107, 326–359.

Chapter 24
Dynamic Policy Analysis

Jaap H. Abbring and James J. Heckman

Abstract This chapter studies the microeconometric treatment-effect and structural approaches to dynamic policy evaluation. First, we discuss a reduced-form approach based on a sequential randomization or dynamic matching assumption that is popular in biostatistics. We then discuss two complementary approaches for treatments that are single stopping times and that allow for non-trivial dynamic selection on unobservables. The first builds on continuous-time duration and event-history models. The second extends the discrete-time dynamic discrete-choice literature.

24.1 Introduction

The methods discussed in Parts 1 and 2 of this volume are useful for microeconometric policy evaluation. That field analyzes the effects of policy interventions on individual outcomes. Panel data facilitate the identification and estimation of such effects. Panel data are especially helpful in analyzing the individual dynamic consequences of policies and outcomes, which are mostly neglected in the vast cross-sectional literature on this topic. Not surprisingly, panel-data methods are becoming more widely used in the microeconometric policy evaluation literature. In this chapter, we critically review recently developed methods and their applications.

The outline of the chapter is as follows. Section 24.2 presents the policy evaluation problem and discusses the treatment-effect approach to policy evaluation. It establishes the notation used in the rest of this chapter. Section 24.3 reviews an approach to the analysis of dynamic treatment effects based on a sequential

Jaap H. Abbring
Department of Economics, VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, and Tinbergen Institute. e-mail: jabbring@econ.vu.nl.

James J. Heckman
Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago IL 60637, USA; American Bar Foundation and Geary Institute, University College Dublin, e-mail: jjh@uchicago.edu.

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

randomization assumption that is popular in biostatistics (Gill and Robins, 2001; Lok 2007; Robins, 1997) and has been applied in economics (see Fitzenberger, Osikominu, and Völter, 2006; and Lechner and Miquel, 2002). This is a dynamic version of matching. We relate the assumptions justifying this approach to the assumptions underlying the econometric dynamic discrete-choice literature based on Rust's (1987) conditional-independence condition which, as discussed in Sect. 24.5.5 below, is frequently invoked in the structural econometrics literature. We note the limitations of the dynamic matching treatment-effect approach in accounting for dynamic information accumulation. In Sects. 24.4 and 24.5, we discuss two econometric approaches for the analysis of treatment times that allow for non-trivial dynamic selection on unobservables. Section 24.4 discusses the continuous-time event-history approach to policy evaluation (Abbring and Van den Berg, 2003b, 2005) and Abbring (2008). Section 24.5 introduces an approach developed by Heckman and Navarro (2007) that builds on and extends the discrete-time dynamic discrete-choice literature. Like the analysis of Abbring and Van den Berg, it does not rely on the conditional-independence assumptions used in dynamic matching. The two complementary approaches surveyed in this chapter span the existing econometric literature on dynamic treatment effects.

24.2 Policy Evaluation and Treatment Effects

24.2.1 The Evaluation Problem

We introduce some key ideas and set up the notation for this chapter by reviewing the static policy evaluation problem discussed in, e.g., Heckman and Vytlacil (2007a). Let Ω be the set of agent types. It is the sample space of a probability space $(\Omega, \mathscr{I}, \mathbb{P})$, and all choices and outcomes are random variables defined on this probability space. Each agent type $\omega \in \Omega$ represents a single agent in a particular state of nature. We could distinguish variation between agents from within-agent randomness by taking $\Omega = J \times \tilde{\Omega}$, with J the set of agents and $\tilde{\Omega}$ the set of possible states of nature. However, we do not make this distinction explicit in this chapter, and often simply refer to agents instead of agent types.[1]

Consider a policy that targets the allocation of each agent in Ω to a single treatment from a set $\mathscr{S}$. In the most basic binary version, $\mathscr{S} = \{0,1\}$, where "1" represents "treatment", such as a training program, and "0" some baseline, "control" program. Alternatively, $\mathscr{S}$ could take a continuum of values, e.g., $\mathbb{R}_+ = [0,\infty)$, representing, e.g., unemployment benefit levels, or duration of time in a program.

A policy $p = (a,\tau) \in \mathscr{A} \times \mathscr{T} \equiv \mathscr{P}$ consists of a planner's rule $a : \Omega \to \mathscr{B}$ for allocating constraints and incentives to agents, and a rule $\tau : \Omega \times \mathscr{A} \to \mathscr{S}$ that generates

[1] For example, we could have $\Omega = [0,1]$ indexing the population of agents, with $\mathbb{P}$ being Lebesgue measure on $[0,1]$. Alternatively, we could take $\Omega = [0,1] \times \tilde{\Omega}$ and have $[0,1]$ represent the population of agents and $\tilde{\Omega}$ states of nature.

agent treatment choices for a given constraint allocation a. This framework allows agent ω's treatment choice to depend both on the constraint assignment mechanism a—in particular, the distribution of the constraints in the population—and on the constraints $a(\omega) \in \mathscr{B}$ assigned to agent ω.

The randomness in the planner's constraint assignment a may reflect heterogeneity of agents as observed by the planner, but it may also be due to explicit randomization. For example, consider profiling on background characteristics of potential participants in the assignment a to treatment eligibility. If the planner observes some background characteristics on individuals in the population of interest, she could choose eligibility status to be a deterministic function of those characteristics and, possibly, some other random variable under her control by randomization. This includes the special case in which the planner randomizes persons into eligibility. We denote the information set generated by the variables observed by the planner when she assigns constraints, including those generated through deliberate randomization, by $\mathscr{I}_P$.[2] The planner's information set $\mathscr{I}_P$ determines how precisely she can target each agent's ω when assigning constraints. The variables in the information set fully determine the constraints assignment a.

Subsequent to the planner's constraints assignment a, each agent ω chooses treatment $\tau(\omega, a)$. We assume that agents know the constraint assignment mechanism a in place. However, agents do not directly observe their types ω, but only observe realizations $I_A(\omega)$ of some random variables I_A. For given $a \in \mathscr{A}$, agent ω's treatment choice $\tau(\omega, a)$ can only depend on ω through his observations $I_A(\omega)$. Typically, $I_A(\omega)$ includes the variables used by the planner in determining $a(\omega)$, so that agents know the constraints that they are facing. Other components of $I_A(\omega)$ may be determinants of preferences and outcomes. Variation in $I_A(\omega)$ across ω may thus reflect preference heterogeneity, heterogeneity in the assigned constraints, and heterogeneity in outcome predictors. We use $\mathscr{I}_A$ to denote the information set generated by I_A.[3] An agent's information set $\mathscr{I}_A$ determines how precisely the agent can tailor his treatment choice to his type ω. For expositional convenience, we assume that agents know more when choosing treatment than what the planner knows when assigning constraints, so that $\mathscr{I}_A \supseteq \mathscr{I}_P$. One consequence is that agents observe the constraints $a(\omega)$ assigned to them, as previously discussed. In turn, the econometrician may not have access to all of the information that is used by the agents when

[2] Formally, $\mathscr{I}_P$ is a sub-σ-algebra of $\mathscr{I}$ and a is assumed to be $\mathscr{I}_P$-measurable.

[3] Formally, $\mathscr{I}_A$ is a sub-σ-algebra of $\mathscr{I}$—the σ-algebra generated by I_A—and $\omega \in \Omega \mapsto \tau(\omega, a) \in \mathscr{S}$ should be $\mathscr{I}_A$-measurable for all $a \in \mathscr{A}$. The possibility that different agents have different information sets is allowed for because a distinction between agents and states of nature is implicit. As suggested in the introduction to this section, we can make it explicit by distinguishing a set J of agents and a set $\tilde{\Omega}$ of states of nature and writing $\Omega = J \times \tilde{\Omega}$. For expositional convenience, let J be finite. We can model the case that agents observe their identity j by assuming that the random variable J_A on Ω that reveals their identity, that is $J_A(j, \tilde{\omega}) = j$, is in their information set $\mathscr{I}_A$. If agents, in addition, observe some other random variable V on Ω, then the information set $\mathscr{I}_A$ generated by (J_A, V) can be interpreted as providing each agent $j \in J$ with perfect information about his identity j and with the agent-j-specific information about the state of nature $\tilde{\omega}$ encoded in the random variable $\tilde{\omega} \mapsto V(j, \tilde{\omega})$ on $\tilde{\Omega}$.

they choose treatment.[4] In this case, $\mathscr{I}_A \not\subseteq \mathscr{I}_E$, where $\mathscr{I}_E$ denotes the econometrician's information set.

We define $s_p(\omega)$ as the treatment selected by agent ω under policy p. With $p = (a, \tau)$, we have that $s_p(\omega) = \tau(\omega, a)$. The random variable $s_p : \Omega \to \mathscr{S}$ represents the allocation of agents to treatments implied by policy p.[5] Randomness in this allocation reflects both heterogeneity in the planner's assignment of constraints and the agents' heterogenous responses to this assignment. One extreme case arises if the planner assigns agents to treatment groups and agents perfectly comply, so that $\mathscr{B} = \mathscr{S}$ and $s_p(\omega) = \tau(\omega, a) = a(\omega)$ for all $\omega \in \Omega$. In this case, all variation of s_p is due to heterogeneity in the constraints $a(\omega)$ across agents ω. At the other extreme, agents do not respond at all to the incentives assigned by mechanisms in $\mathscr{A}$, and $\tau(a, \omega) = \tau(a', \omega)$ for all $a, a' \in \mathscr{A}$ and $\omega \in \Omega$. In general, there are policies that have a nontrivial (that is, nondegenerate) constraint assignment a, where at least some agents respond to the assigned constraints a in their treatment choice, $\tau(a, \omega) \neq \tau(a', \omega)$ for some $a, a' \in \mathscr{A}$ and $\omega \in \Omega$.

We seek to evaluate a policy p in terms of some outcome Y_p, for example, earnings. For each $p \in \mathscr{P}$, Y_p is a random variable defined on the population Ω. The evaluation can focus on objective outcomes Y_p, on the subjective valuation $R(Y_p)$ of Y_p by the planner or the agents, or on both types of outcomes. The evaluation can be performed relative to a variety of information sets reflecting different actors (the agent, the planner and the econometrician) and the arrival of information in different time periods. Thus, the randomness of Y_p may represent both (*ex ante*) heterogeneity among agents known to the planner when constraints are assigned (that is, variables in $\mathscr{I}_P$) and/or heterogeneity known to the agents when they choose treatment (that is, information in $\mathscr{I}_A$), as well as (*ex post*) shocks that are not foreseen by the policy maker or by the agents. An information-feasible (*ex ante*) policy evaluation by the planner would be based on some criterion using the distribution of Y_p conditional on $\mathscr{I}_P$. The econometrician can assist the planner in computing this evaluation if the planner shares her *ex ante* information and $\mathscr{I}_P \subseteq \mathscr{I}_E$.

Suppose that we have data on outcomes Y_{p_0} under policy p_0 with corresponding treatment assignment s_{p_0}. Consider an intervention that changes the policy from the actual p_0 to some counterfactual p' with associated treatments $s_{p'}$ and outcomes $Y_{p'}$. This could involve a change in the planner's constraint assignment from a_0 to a' for given $\tau_0 = \tau'$, a change in the agent choice rule from τ_0 to τ' for given $a_0 = a'$, or both.

The policy evaluation problem involves contrasting $Y_{p'}$ and Y_{p_0} or functions of these outcomes. For example, if the outcome of interest is mean earnings, we might be interested in some weighted average of $E[Y_{p'} - Y_{p_0} \mid \mathscr{I}_P]$, such as $E[Y_{p'} - Y_{p_0}]$. The special case where $\mathscr{S} = \{0, 1\}$ and $s_{p'} = a' = 0$ generates the effect of abolishing the program. Implementing such a policy requires that the planner be able to induce all agents into the control group by assigning constraints $a' = 0$. In particular, this

[4] See the discussion by Heckman and Vytlacil (2007b, Sects. 2 and 9).

[5] Formally, $\{s_p\}_{p \in \mathscr{A} \times \mathscr{T}}$ is a stochastic process indexed by p.

assumes that there are no substitute programs available to agents that are outside the planner's control (Heckman and Vytlacil, 2007b, Sect. 10).

For notational convenience, write $S = s_{p_0}$ for treatment assignment under the actual policy p_0 in place. Cross-sectional micro data typically provide a random sample from the joint distribution of (Y_{p_0}, S).[6] Clearly, without further assumptions, such data do not identify the effects of the policy shift from p_0 to p'. This identification problem becomes even more difficult if we do not seek to compare the counterfactual policy p' with the actual policy p_0, but rather with another counterfactual policy p'' that also has never been observed. A leading example is the binary case in which $0 < \Pr(S = 1) < 1$, but we seek to know the effects of $s_{p'} = 0$ (universal nonparticipation) and $s_{p''} = 1$ (universal treatment), where neither policy has ever been observed in place.

Panel data can help to evaluate the type of static policies discussed so far, if interpreted as short-run or even one-shot policies. Suppose that we have data on outcomes in two periods in which two different policies were in place. In a world in which outcomes in any period are not affected by the policy or outcomes in the other period, such data are directly informative on the contrast between outcomes under both policies.

The standard microeconometric approach to the policy evaluation problem assumes that the (subjective and objective) outcomes for any individual agent are the same across all policy regimes for any particular treatment assigned to the individual (see, e.g., Heckman, LaLonde, and Smith, 1999). Heckman and Vytlacil (2007a) present a detailed account of the policy-invariance assumptions that justify this practice. They simplify the task of evaluating policy p to determining (i) the assignment s_p of treatments under policy p and (ii) treatment effects for individual outcomes. Even within this simplified framework, there are still two difficult, and distinct, problems in identifying treatment effects on individual outcomes:

(A) *The Evaluation Problem: that we observe an agent in one treatment state and seek to determine that agent's outcomes in another state; and*
(B) *The Selection Problem: that the distributions of outcomes for the agents we observe in a given treatment state are not the marginal population distributions that would be observed if agents were randomly assigned to the state.*

The assignment mechanism s_p of treatments under counterfactual policies p is straightforward in the case where the planner assigns agents to treatment groups and agents fully comply, so that $s_p = a$. More generally, an explicit model of agent treatment choices is needed to derive s_p for counterfactual policies p. An explicit model of agent treatment choices can also be helpful in addressing the selection problem, and in identifying agent subjective valuations of outcomes. We now formalize the notation for the treatment-effect approach that we will use in this section using the

[6] Notice that a random sample of outcomes under a policy may entail nonrandom selection of treatments as individual agents select individual treatments given τ and the constraints they face assigned by a.

potential-outcome framework of Neyman (1923), Roy (1951), Quandt (1958, 1972), and Rubin (1974).[7]

24.2.2 The Treatment Effect Approach

For each agent $\omega \in \Omega$, let $y(s, X(\omega), U(\omega))$ be the potential outcome when the agent is assigned to treatment $s \in \mathscr{S}$. Here, X and U are covariates that are not causally affected by the treatment or the outcomes.[8,9] In the language of Kalbfleisch and Prentice (1980) and Leamer (1985), we say that such covariates are "external" to the causal model. X is observed by the econometrician (that is, in $\mathscr{I}_E$) and U is not.

Recall that s_p is the assignment of agents to treatments under policy p. For all policies p that we consider, the outcome Y_p is linked to the potential outcomes by the consistency condition $Y_p = y(s_p, X, U)$. This condition follows from the policy-invariance assumptions. It embodies the assumption that an agent's outcome only depends on the treatment assigned to the agent and not separately on the mechanism used to assign treatments. This excludes (strategic) interactions between agents and equilibrium effects of the policy.[10] It ensures that we can specify individual outcomes y from participating in programs in $\mathscr{S}$ independently of the policy p and treatment assignment s_p. Economists say that y is autonomous, or structurally invariant with respect to the policy environment (see Frisch, 1938; Hurwicz, 1962; and Heckman and Vytlacil, 2007a).[11]

We illustrate the treatment-effect approach with a basic example. Consider the evaluation of an intervention that changes the policy from p_0 to p' in terms of its mean effect $E[Y_{p'} - Y_{p_0}]$ on outcomes. For expositional convenience, let treatment be binary: $\mathscr{S} = \{0, 1\}$. Suppose that we have a cross-sectional sample from the joint distribution of (Y_{p_0}, S, X). Assume that treatment assignment under both the actual policy p_0 and the alternative policy p' is randomized, that is, both S and $s_{p'}$ are independent of the determinants (X, U) of the potential outcomes. Then, because of the policy-invariance conditions,

---

[7] See Heckman et al. (1999); Heckman and Vytlacil (2007a); and Thurstone (1930), for results in econometrics and extensive reviews of the econometric literature.

[8] This "no feedback" condition requires that X and U are the same fixing treatment to s for all s. See Haavelmo (1943), Pearl (2000), or the discussion in Heckman and Vytlacil (2007a,b).

[9] Note that this framework is rich enough to capture the case in which potential outcomes depend on treatment-specific unobservables as in Sect. 24.5, because these can be simply stacked in U and subsequently selected by y. For example, in the case where $\mathscr{S} = \{0, 1\}$ we can write $y(s, X, (U_0, U_1)) = s y_1(X, U_1) + (1 - s) y_0(X, U_0)$ for some y_0 and y_1. A specification without treatment-dependent unobservables is more tractable in the case of continuous treatments in Sect. 24.3 and, in particular, continuous treatment times in Sect. 24.4.

[10] See Pearl (2000); Heckman (2005); and Heckman and Vytlacil (2007a).

[11] See also Aldrich (1989) and Hendry and Morgan (1995). Rubin's (1986) stable-unit-treatment-value assumption is a version of the classical invariance assumptions of econometrics (see Abbring, 2003; and Heckman and Vytlacil, 2007a, for discussion of this point).

$$E[Y_{p'} - Y_{p_0}] = E[y(1,X,U) - y(0,X,U)] \left[\Pr(s_{p'} = 1) - \Pr(S = 1)\right].$$

The mean effect of the intervention on outcomes equals the "average treatment effect" $E[y(1,X,U) - y(0,X,U)]$ times the net increase in the assignment to treatment 1. The policy evaluation problem boils down to identifying the average treatment effect, the distribution of the actual treatment assignment S, and the distribution of treatment assignment $s_{p'}$ under the alternative policy p'. Under the assumption of randomized assignment, and provided that $0 < \Pr(S = 1) < 1$, the average treatment effect is identified as $E[Y_{p_0}|S = 1] - E[Y_{p_0}|S = 0]$. The distribution of S is identified directly from the data. The distribution of $s_{p'}$ is often known, as in the case of universal nonparticipation ($s_{p'} = 0$) or universal treatment ($s_{p'} = 1$). Otherwise, it needs to be identified using a model of treatment choice.

Heckman and Vytlacil (2007a,b) review more general evaluation problems and econometric methods that do not rely on randomized assignment, such as the methods of matching and instrumental variables. Clearly, panel data, combined with stationarity assumptions, can help in addressing the selection problem in the evaluation of static policies. We will not dwell on this application of panel data to the evaluation of static policies, but now turn to the dynamic policy evaluation problem.

24.2.3 Dynamic Policy Evaluation

Interventions often have consequences that span over many periods. Policy interventions at different points in time can be expected to affect not only current outcomes, but also outcomes at other points in time. The same policy implemented at different time periods may have different consequences. Moreover, policy assignment rules often have non-trivial dynamics. The assignment of programs at any point in time can be contingent on the available data on past program participation, intermediate outcomes and covariates.

The dynamic policy evaluation problem can be formalized in a fashion similar to the way we formalized the static problem in Sect. 24.2.1. In this subsection, we analyze a discrete-time finite-horizon model. We consider continuous-time models in Sect. 24.4. The possible treatment assignment times are $1,\ldots,\bar{T}$. We do not restrict the set $\mathscr{S}$ of treatments. We allow the same treatment to be assigned on multiple occasions. In general, the set of available treatments at each time t may depend on time t and on the history of treatments, outcomes, and covariates. For expositional convenience, we will only make this explicit in Sects. 24.4 and 24.5, where we focus on the timing of a single treatment.

We define a dynamic policy $p = (a,\tau) \in \mathscr{A} \times \mathscr{T} \equiv \mathscr{P}$ as a dynamic constraint assignment rule $a = \{a_t\}_{t=1}^{\bar{T}}$ with a dynamic treatment choice rule $\tau = \{\tau_t\}_{t=1}^{\bar{T}}$. At each time t, the planner assigns constraints $a_t(\omega)$ to each agent $\omega \in \Omega$, using information in the time-t policy-p information set $\mathscr{I}_P(t,p) \subseteq \mathscr{I}$. The planner's information set $\mathscr{I}_P(t,p)$ could be based on covariates and random variables under the planner's control, as well as past choices and realized outcomes. We denote the sequence of

planner's information sets by $\mathscr{I}_P(p) = \{\mathscr{I}_P(t,p)\}_{t=1}^{\bar{T}}$. We assume that the planner does not forget any information she once had, so that her information improves over time and $\mathscr{I}_P(t,p) \subseteq \mathscr{I}_P(t+1,p)$ for all t.[12]

Each agent ω chooses treatment $\tau_t(\omega,a)$ given their information about ω at time t under policy p and given the constraint assignment mechanism $a \in \mathscr{A}$ in place. We assume that agents know the constraint assignment mechanism a in place. At time t, under policy p, agents infer their information about their type ω from random variables $I_A(t,p)$ that may include preference components and determinants of constraints and future outcomes. $\mathscr{I}_A(t,p)$ denotes the time-t policy-p information set generated by $I_A(t,p)$ and $\mathscr{I}_A(p) = \{\mathscr{I}_A(t,p)\}_{t=1}^{\bar{T}}$. We assume that agents are increasingly informed as time goes by, so that $\mathscr{I}_A(t,p) \subseteq \mathscr{I}_A(t+1,p)$.[13] For expositional convenience, we also assume that agents know more than the planner at each time t, so that $\mathscr{I}_P(t,p) \subseteq \mathscr{I}_A(t,p)$.[14] Because all determinants of past and current constraints are in the planner's information set $\mathscr{I}_P(t,p)$, this implies that agents observe $(a_1(\omega),\ldots,a_t(\omega))$ at time t. Usually, they do not observe all determinants of their future constraints $(a_{t+1}(\omega),\ldots,a_{\bar{T}}(\omega))$.[15] Thus, the treatment choices of the agents may be contingent on past and current constraints, their preferences, and on their predictions of future outcomes and constraints given their information $\mathscr{I}_A(t,p)$ and given the constraint assignment mechanism a in place.

Extending the notation for the static case, we denote the assignment of agents to treatment τ_t at time t implied by a policy p by the random variable $s_p(t)$ defined so that $s_p(\omega,t) = \tau_t(\omega,a)$. We use the shorthand s_p^t for the vector $(s_p(1),\ldots,s_p(t))$ of treatments assigned up to and including time t under policy p, and write $s_p = s_p^{\bar{T}}$. The assumptions made so far about the arrival of information imply that treatment assignment $s_p(t)$ can only depend on the information $\mathscr{I}_A(t,p)$ available to agents at time t.[16]

Because past outcomes typically depend on the policy p, the planner's information $\mathscr{I}_P(p)$ and the agents' information $\mathscr{I}_A(p)$ will generally depend on p as well. In the treatment-effect framework that we develop in the next section, at each time t different policies may have selected different elements in the set of potential outcomes in the past. The different elements reveal different aspects of the unobservables underlying past and future outcomes. We will make assumptions that limit the dependence of information sets on policies in the context of the treatment-effect approach developed in the next section.

Objective outcomes associated with policies p are expressed as a vector of time-specific outcomes $Y_p = (Y_p(1),\ldots,Y_p(\bar{T}))$. The components of this vector may also be vectors. We denote the outcomes from time 1 to time t under policy p by

[12] Formally, the information $\mathscr{I}_P(p)$ that accumulates for the planner under policy p is a filtration in $\mathscr{I}$, and a is a stochastic process that is adapted to $\mathscr{I}_P(p)$.

[13] Formally, the information $\mathscr{I}_A(p)$ that accumulates for the agents is a filtration in $\mathscr{I}$.

[14] If agents are strictly better informed, and $\mathscr{I}_P(t,p) \subset \mathscr{I}_A(t,p)$, it is unlikely that the planner catches up and learns the agent's information with a delay (e.g., $\mathscr{I}_A(t,p) \subseteq \mathscr{I}_P(t+1,p)$) unless agent's choices and outcomes reveal all their private information.

[15] Formally, $a_1,\ldots,a_t$ are $\mathscr{I}_A(t,p)$-measurable, but $a_{t+1},\ldots,a_{\bar{T}}$ are not.

[16] Formally, $\{s_p(t)\}_{t=1}^{\bar{T}}$ is a stochastic process that is adapted to $\mathscr{I}_A(p)$.

$Y_p^t = (Y_p(1), \ldots, Y_p(t))$. We analyze both subjective and objective evaluations of policies in Sect. 24.5, where we consider more explicit economic models. Analogous to our analysis of the static case, we cannot learn about the outcomes $Y_{p'}$ that would arise under a counterfactual policy p' from data on outcomes Y_{p_0} and treatments $s_{p_0} = S$ under a policy $p_0 \neq p'$ without imposing further structure on the problem.[17] We follow the approach exposited for the static case and assume policy invariance of individual outcomes under a given treatment. This reduces the evaluation of a dynamic policy p to identifying (i) the dynamic assignment s_p of treatments under policy p and (ii) the dynamic treatment effects on individual outcomes. We focus our discussion on the fundamental evaluation problem and the selection problem that haunt inference about treatment effects. In the remainder of the section, we review alternative approaches to identifying dynamic treatment effects, and some approaches to modeling dynamic treatment choice. We first analyze methods recently developed in statistics.

24.3 Dynamic Treatment Effects and Sequential Randomization

In a series of papers, Robins extends the static Neyman–Roy–Rubin model based on selection on observables to a dynamic setting (see, e.g., Robins, 1997, and the references therein). He does not consider agent choice or subjective evaluations. Here, we review his extension, discuss its relationship to dynamic choice models in econometrics, and assess its merits as a framework for economic policy analysis. We follow the exposition of Gill and Robins (2001), but add some additional structure to their basic framework to exposit the connection of their approach to the dynamic approach pursued in econometrics.

24.3.1 Dynamic Treatment Effects

24.3.1.1 Dynamic Treatment and Dynamic Outcomes

To simplify the exposition, suppose that $\mathscr{S}$ is a finite discrete set.[18] Recall that, at each time t and for given p, treatment assignment $s_p(t)$ is a random variable that only depends on the agent's information $\mathscr{I}_A(t,p)$, which includes personal knowledge of preferences and determinants of constraints and outcomes. To make this dependence explicit, suppose that external covariates Z, observed by the econometrician (that is, variables in $\mathscr{I}_E$), and unobserved external covariates V_1 that

[17] If outcomes under different policy regimes are informative about the same technology and preferences, for example, then the analyst and the agent could learn about the ingredients that produce counterfactual outcomes in all outcome states.

[18] All of the results presented in this subsection extend to the case of continuous treatments. We will give references to the appropriate literature in subsequent footnotes.

affect treatment assignment are revealed to the agents at time 1. Then, at the start of each period $t \geq 2$, past outcomes $Y_p(t-1)$ corresponding to the outcomes realized under treatment assignment s_p and external unobserved covariates V_t enter the agent's information set.[19] In this notation, $\mathscr{I}_A(1,p)$ is the information $\sigma(Z,V_1)$ conveyed to the agent by (Z,V_1) and, for $t \geq 2$, $\mathscr{I}_A(t,p) = \sigma(Y_p^{t-1},Z,V^t)$, with $V^t = (V_1,\ldots,V_t)$. In the notation of the previous subsection, $I_A(1,p) = (Z,V_1)$ and, for $t \geq 2$, $I_A(t,p) = (Y_p^{t-1},Z,V^t)$. Among the elements of $I_A(t,p)$ are the determinants of the constraints faced by the agent up to t, which may or may not be observed by the econometrician.

We attach *ex post* potential outcomes $Y(t,s) = y_t(s,X,U_t)$, $t = 1,\ldots,\bar{T}$, to each treatment sequence $s = (s(1),\ldots,s(\bar{T}))$. Here, X is a vector of observed (by the econometrician) external covariates and U_t, $t = 1,\ldots,\bar{T}$, are vectors of unobserved external covariates. Some components of X and U_t may be in agent information sets. We denote $Y^t(s) = (Y(1,s),\ldots,Y(t,s))$, $Y(s) = Y^{\bar{T}}(s)$, and $U = (U_1,\ldots,U_{\bar{T}})$. As in the static case, potential outcomes y are assumed to be invariant across policies p, which ensures that $Y_p(t) = y_t(s_p,X,U_t)$. In the remainder of this section, we keep the dependence of outcomes on observed covariates X implicit and suppress all conditioning on X.

We assume no causal dependence of outcomes on future treatment:[20]

(NA) *For all* $t \geq 1$, $Y(t,s) = Y(t,s')$ *for all* s,s' *such that* $s^t = (s')^t$,

where $s^t = (s(1),\ldots,s(t))$ and $(s')^t = (s'(1),\ldots,s'(t))$. Abbring and Van den Berg (2003b) and Abbring (2003) define this as a "no-anticipation" condition. It requires that outcomes at time t (and before) be the same across policies that allocate the same treatment up to and including t, even if they allocate different treatments after t. In the structural econometric models discussed in Sects. 24.3.2.2 and 24.5 below, this condition is trivially satisfied if all state variables relevant to outcomes at time t are included as inputs in the outcome equations $Y(t,s) = y_t(s,U_t)$, $t = 1,\ldots,\bar{T}$.

Because Z and V_1 are assumed to be externally determined, and therefore not affected by the policy p, the initial agent information set $\mathscr{I}_A(1,p) = \sigma(Z,V_1)$ does not depend on p. Agent ω has the same initial data $(Z(\omega),V_1(\omega))$ about his type ω under all policies p. Thus, $\mathscr{I}_A(1,p) = \mathscr{I}_A(1,p')$ is a natural benchmark information set for an *ex ante* comparison of outcomes at time 1 among different policies. For $t \geq 2$, (NA) implies that actual outcomes up to time $t-1$ are equal between policies p and p', $Y_p^{t-1} = Y_{p'}^{t-1}$, if the treatment histories coincide up to time $t-1$ so that $s_p^{t-1} = s_{p'}^{t-1}$. Together with the assumption that Z and V^t are externally determined, it follows that agents have the same time-t information set structure about ω under policies p and p', $\mathscr{I}_A(t,p) = \sigma(Y_p^{t-1},Z,V^t) = \sigma(Y_{p'}^{t-1},Z,V^t) = \mathscr{I}_A(t,p')$,

[19] Note that any observed covariates that are dynamically revealed to the agents can be subsumed in the outcomes.

[20] For statistical inference from data on the distribution of (Y_{p_0},S,Z), these equalities only need to hold on events $\{\omega \in \Omega : S^t(\omega) = s^t\}$, $t \geq 1$, respectively.

if $s_p^{t-1} = s_{p'}^{t-1}$.[21,22] In this context, $\mathscr{I}_A(t,p) = \mathscr{I}_A(t,p')$ is a natural information set for an *ex ante* comparison of outcomes from time t onwards between any two policies p and p' such that $s_p^{t-1} = s_{p'}^{t-1}$.

With this structure on the agent information sets in hand, it is instructive to review the separate roles in determining treatment choice of information about ω and knowledge about the constraint assignment rule a. First, agent ω's time-t treatment choice $s_p(\omega,t) = \tau_t(\omega,a)$ may depend on distributional properties of a, for example the share of agents assigned to particular treatment sequences, and on the past and current constraints $(a_1(\omega),\ldots,a_t(\omega))$ that were actually assigned to him. We have assumed both to be known to the agent. Both may differ between policies, even if the agent information about ω is fixed across the policies. Second, agent ω's time-t treatment choice may depend on agent ω's predictions of future constraints and outcomes. A forward-looking agent ω will use observations of his covariates $Z(\omega)$ and $V^t(\omega)$ and past outcomes $Y_p^{t-1}(\omega)$ to infer his type ω and subsequently predict future external determinants $(U_t(\omega),\ldots,U_{\bar{T}}(\omega))$ of his outcomes and $(V_{t+1}(\omega),\ldots,V_{\bar{T}}(\omega))$ of his constraints and treatments. In turn, this information updating allows agent ω to predict his future potential outcomes $(Y(t,s,\omega),\ldots,Y(\bar{T},s,\omega))$ and, for a given policy regime p, his future constraints $(a_{t+1}(\omega),\ldots,a_{\bar{T}}(\omega))$, treatments $(s_p(t+1,\omega),\ldots,s_p(\bar{T},\omega))$, and realized outcomes $(Y_p(t,\omega),\ldots,Y_p(\bar{T},\omega))$. Under different policies, the agent may gather different information on his type ω and therefore come up with different predictions of the external determinants of his future potential outcomes and constraints. In addition, even if the agent has the same time-t predictions of the external determinants of future constraints and potential outcomes, he may translate these into different predictions of future constraints and outcomes under different policies.

Assumption (NA) requires that current potential outcomes are not affected by future treatment. Justifying this assumption requires specification of agent information about future treatment and agent behavior in response to that information. Such an interpretation requires that we formalize how information accumulates for agents across treatment sequences s and s' such that $s^t = (s')^t$ and $(s_{t+1},\ldots,s_{\bar{T}}) \neq (s'_{t+1},\ldots,s'_{\bar{T}})$. To this end, consider policies p and p' such that $s_p = s$ and $s_{p'} = s'$. These policies produce the same treatment assignment up to time t, but are different in the future. We have previously shown that, even though the time-t agent information about ω is the same under both policies, $\mathscr{I}_A(t,p) = \mathscr{I}_A(t,p')$, agents may have different predictions of future constraints, treatments and outcomes because the policies may differ in the future and agents know this. The policy-invariance

[21] If $s_p^{t-1}(\omega) = s_{p'}^{t-1}(\omega)$ only holds for ω in some subset $\Omega_{t-1} \subset \Omega$ of agents, then $Y_p^{t-1}(\omega) = Y_{p'}^{t-1}(\omega)$ only for $\omega \in \Omega_{t-1}$, and information coincides between p and p' only for agents in Ω_{t-1}. Formally, let Ω_{t-1} be the set $\{\omega \in \Omega : s_p^{t-1}(\omega) = s_{p'}^{t-1}(\omega)\}$ of agents that share the same treatment up to and including time $t-1$. Then, Ω_{t-1} is in the agent's information set under both policies, $\Omega_{t-1} \in \mathscr{I}_A(t,p) \cap \mathscr{I}_A(t,p')$. Moreover, the partitioning of Ω_{t-1} implied by $\mathscr{I}_A(t,p)$ and $\mathscr{I}_A(t,p')$ is the same. To see this, note that the collections of all sets in, respectively, $\mathscr{I}_A(t,p)$ and $\mathscr{I}_A(t,p')$ that are weakly included in Ω_{t-1} are identical σ-algebras on Ω_{t-1}.

[22] Notice that the realizations of the random variables $Y_{p'}^{t-1}$, Z, V^t may differ among agents.

conditions ensure that time-t potential outcomes are nevertheless the same under each policy. This requires that potential outcomes be determined externally, and are not affected by agent actions in response to different predictions of future constraints, treatments and outcomes. This assumption rules out investment responses to alternative policies that affect potential outcomes.

In general, different policies in $\mathscr{P}$ will produce different predictions of future constraints, treatment and outcomes. In the dynamic treatment-effect framework, this may affect outcomes indirectly through agent treatment choices. If potential outcomes are directly affected by agent's forward-looking decisions, then the invariance conditions underlying the treatment-effect framework will be violated. Section 24.3.3 illustrates this issue, and the no-anticipation condition, with some examples.

24.3.1.2 Identification of Treatment Effects

Suppose that the econometrician has data that allows her to estimate the joint distribution of (Y_{p_0}, S, Z) of outcomes, treatments and covariates under some policy p_0, where $S = s_{p_0}$. These data are not enough to identify dynamic treatment effects.

To secure identification, Gill and Robins (2001) invoke a dynamic version of the matching assumption (conditional independence) which relies on sequential randomization:[23]

(M-1) *For all treatment sequences s and all t,*

$$S(t) \perp\!\!\!\perp (Y(t,s), \ldots, Y(\bar{T},s)) \mid (Y_{p_0}^{t-1}, S^{t-1} = s^{t-1}, Z),$$

where the conditioning set $(Y_{p_0}^0, S^0 = s^0, Z)$ for $t = 1$ is Z.

Equivalently, $S(t) \perp\!\!\!\perp (U_t, \ldots, U_{\bar{T}}) \mid (Y_{p_0}^{t-1}, S^{t-1}, Z)$ for all t without further restricting the data. Sequential randomization allows the $Y_{p_0}(t)$ to be "dynamic confounders"—variables that are affected by past treatment and that affect future treatment assignment.

The sequence of conditioning information sets appearing in the sequential randomization assumption, $\mathscr{I}_E(1) = \sigma(Z)$ and, for $t \geq 2$, $\mathscr{I}_E(t) = \sigma(Y_{p_0}^{t-1}, S^{t-1}, Z)$, is a filtration $\mathscr{I}_E$ of the econometrician's information set $\sigma(Y_{p_0}, S, Z)$. Note that $\mathscr{I}_E(t) \subseteq \mathscr{I}_A(t, p_0)$ for each t. If treatment assignment is based on strictly more information than $\mathscr{I}_E$, so that agents know strictly more than the econometrician and act on their superior information, (M-1) is likely to fail if that extra information also affects outcomes. Heckman and Vytlacil (2007b) make this point in a static setting.

[23] Formally, we need to restrict attention to sequences s in the support of S. Throughout this section, we will assume this and related support conditions hold.

Together with the no-anticipation condition (NA), which is a condition on outcomes and distinct from (M-1), the dynamic potential-outcome model set up so far is a natural dynamic extension of the Neyman–Roy–Rubin model for a static (stratified) randomized experiment.

Under assumption (M-1) that the actual treatment assignment S is sequentially randomized, we can sequentially identify the causal effects of treatment from the distribution of the data (Y_{p_0}, S, Z) and construct the distribution of the potential outcomes $Y(s)$ for any treatment sequence s in the support of S.

Consider the case in which all variables are discrete. No-anticipation condition (NA) ensures that potential outcomes for a treatment sequence s equal actual (under policy p_0) outcomes up to time $t-1$ for agents with treatment history s^{t-1} up to time $t-1$. Formally, $Y^{t-1}(s) = Y_{p_0}^{t-1}$ on the set $\{S^{t-1} = s^{t-1}\}$. Using this, sequential randomization assumption (M-1) can be rephrased in terms of potential outcomes: for all s and t,

$$S(t) \perp\!\!\!\perp (Y(t,s), \ldots, Y(\bar{T},s)) \mid (Y^{t-1}(s), S^{t-1} = s^{t-1}, Z) \,.$$

In turn, this implies that, for all s and t,

$$\begin{aligned} \Pr\left(Y(t,s) = y(t) \mid Y^{t-1}(s) = y^{t-1}, S^t = s^t, Z\right) \\ = \Pr\left(Y(t,s) = y(t) \mid Y^{t-1}(s) = y^{t-1}, Z\right) \,, \end{aligned} \tag{24.1}$$

where $y^{t-1} = (y(1), \ldots, y(t-1))$ and $y = y^{\bar{T}}$. From Bayes' rule and (24.1), it follows that

$$\begin{aligned} &\Pr(Y(s) = y|Z) \\ &= \Pr(Y(1,s) = y(1) \mid Z) \prod_{t=2}^{\bar{T}} \Pr\left(Y(t,s) = y(t) \mid Y^{t-1}(s) = y^{t-1}, Z\right) \\ &= \Pr(Y(1,s) = y(1) \mid S(1) = s(1), Z) \\ &\quad \times \prod_{t=2}^{\bar{T}} \Pr\left(Y(t,s) = y(t) \mid Y^{t-1}(s) = y^{t-1}, S^t = s^t, Z\right) \,. \end{aligned}$$

Invoking (NA), in particular $Y(t,s) = Y_{p_0}(t)$ and $Y^{t-1}(s) = Y_{p_0}^{t-1}$ on $\{S^t = s^t\}$, produces

$$\begin{aligned} \Pr(Y(s) = y|Z) = \Pr\left(Y_{p_0}(1) = y(1) \mid S(1) = s(1), Z\right) \\ \times \prod_{t=2}^{\bar{T}} \Pr\left(Y_{p_0}(t) = y(t) \mid Y_{p_0}^{t-1} = y^{t-1}, S^t = s^t, Z\right) \,. \end{aligned} \tag{24.2}$$

This is a version of Robins' (1997) "g-computation formula".[24,25] We can sequentially identify each component on the left hand side of the first expression, and hence identify the counterfactual distributions. This establishes identification of the distribution of $Y(s)$ by expressing it in terms of objects that can be identified from data. Identification is exact (or "tight") in the sense that the identifying assumptions, no anticipation and sequential randomization, do not restrict the factual data and are therefore not testable (Gill and Robins, 2001).[26]

Example 24.1. Consider a two-period ($\bar{T} = 2$) version of the model in which agents take either "treatment" (1) or "control" (0) in each period. Then, $S(1)$ and $S(2)$ have values in $\mathscr{S} = \{0,1\}$. The potential outcomes in period t are $Y(t,(0,0))$, $Y(t,(0,1))$, $Y(t,(1,0))$ and $Y(t,(1,1))$. For example, $Y(2,(0,0))$ is the outcome in period 2 in the case that the agent is assigned to the control group in each of the two periods. Using Bayes' rule, it follows that

$$\Pr(Y(s) = y|Z) = \Pr(Y(1,s) = y(1) \mid Z) \Pr(Y(2,s) = y(2) \mid Y(1,s) = y(1), Z) \ . \tag{24.3}$$

The g-computation approach to constructing $\Pr(Y(s) = y|Z)$ from data replaces the two probabilities in the right-hand side with probabilities of the observed (by the econometrician) variables (Y_{p_0}, S, Z). First, note that $\Pr(Y(1,s) = y(1) \mid Z) = \Pr(Y(1,s) = y(1) \mid S(1) = s(1), Z)$ by (M-1). Moreover, (NA) ensures that potential outcomes in period 1 do not depend on the treatment status in period 2, so that

$$\Pr(Y(1,s) = y(1) \mid Z) = \Pr\left(Y_{p_{(0)}}(1) = y(1) \mid S(1) = s(1), Z\right) \ .$$

[24] Gill and Robins (2001) present versions of (NA) and (M-1) for the case with more general distributions of treatments, and prove a version of the g-computation formula for the general case. For a random vector X and a function f that is integrable with respect to the distribution of X, let $\int_{x\in A} f(x) \Pr(X \in \mathrm{d}x) = E[f(X)\mathbf{1}(X \in A)]$. Then,

$$\begin{aligned} \Pr(Y(s) \in A|Z) = \int_{y\in A} & \Pr\left(Y_{p_0}(\bar{T}) \in \mathrm{d}y(\bar{T}) \mid Y_{p_0}^{\bar{T}-1} = y^{\bar{T}-1}, S^{\bar{T}} = s^{\bar{T}}, Z\right) \\ & \vdots \\ & \times \Pr\left(Y_{p_0}(2) \in \mathrm{d}y(2) \mid Y_{p_0}(1) = y(1), S^2 = s^2, Z\right) \\ & \times \Pr\left(Y_{p_0}(1) \in \mathrm{d}y(1) \mid S(1) = s(1), Z\right) \ , \end{aligned}$$

where A is a set of $Y(s)$. The right-hand side of this expression is almost surely unique under regularity conditions presented by Gill and Robins (2001).

[25] An interesting special case arises if the outcomes are survival indicators, that is if $Y_{p_0}(t) = 1$ if the agent survives up to and including time t and $Y_{p_0}(t) = 0$ otherwise, $t \geq 1$. Then, no anticipation (NA) requires that treatment after death does not affect survival, and the g-computation formula simplifies considerably (Abbring, 2003).

[26] Gill and Robins' (2001) analysis only involves causal inference on a final outcome $\left(\text{i.e., our } Y\left(s,\bar{T}\right)\right)$ and does not invoke the no-anticipation condition. However, their proof directly applies to the case studied in this chapter.

Similarly, subsequently invoking (NA) and (M-1), then (M-1), and then (NA), gives

$$\begin{aligned}
&\Pr\left(Y(2,s)=y(2)\mid Y(1,s)=y(1),Z\right)\\
&=\Pr\left(Y(2,s)=y(2)\mid Y_{p_0}(1),S(1)=s(1),Z\right) \qquad \text{(by (NA) and (M-1))}\\
&=\Pr\left(Y(2,s)=y(2)\mid Y_{p_0}(1),S=s,Z\right) \qquad \text{(by (M-1))}\\
&=\Pr\left(Y_{p_0}(2)=y(2)\mid Y_{p_0}(1),S=s,Z\right)\ . \qquad \text{(by (NA))}
\end{aligned}$$

Substituting these equations into the right-hand side of (24.3) gives the g-computation formula,

$$\begin{aligned}
\Pr\left(Y(s)\ =\ y|Z\right)&=\Pr\left(Y_{p_0}(1)=y(1)\mid S(1)=s(1),Z\right)\\
&\quad\times\Pr\left(Y_{p_0}(2)\ =\ y(2)\mid Y_{p_0}(1)=y(1),S=s,Z\right)\ .
\end{aligned}$$

Note that the right-hand side expression does not generally reduce to $\Pr\left(Y_{p_0}=y|\right.$ $\left.S=s,Z\right)$. This would require the stronger, static matching condition $S\perp\!\!\!\perp Y(s)\mid Z$, which we have not assumed here.

Matching on pretreatment covariates is a special case of the g-computation approach. Suppose that the entire treatment path is assigned independently of potential outcomes given pretreatment covariates Z or, more precisely, $S\perp\!\!\!\perp Y(s)\mid Z$ for all s. This implies sequential randomization (M-1), and directly gives identification of the distributions of $Y(s)|Z$ and $Y(s)$. The matching assumption imposes no restriction on the data since $Y(s)$ is only observed if $S=s$. The no-anticipation condition (NA) is not required for identification in this special case because no conditioning on S^t is required. Matching on pretreatment covariates is equivalent to matching in a static model. The distribution of $Y(s)|Z$ is identified without (NA), and assuming it to be true would impose testable restrictions on the data. In particular, it would imply that treatment assignment cannot be dependent on past outcomes given Z. The static matching assumption is not likely to hold in applications where treatment is dynamically assigned based on information on intermediate outcomes. This motivates an analysis based on the more subtle sequential randomization assumption. An alternative approach, developed in Sect. 24.5, is to explicitly model and identify the evolution of the unobservables.

Gill and Robins claim that their sequential randomization and no-anticipation assumptions are "neutral", "for free", or "harmless". As we will argue later, from an economic perspective, some of the model assumptions, notably the no-anticipation assumption, can be interpreted as substantial behavioral/informational assumptions. For example, Heckman and Vytlacil (2005, 2007b) and Heckman and Navarro (2004) show how matching imposes the condition that marginal and average returns are equal. Because of these strong assumptions, econometricians sometimes phrase their "neutrality" result more negatively as a non-identification result (Abbring and Van den Berg, 2003b), since it is possible that (M-1) and/or (NA) may not hold.

24.3.2 Policy Evaluation and Dynamic Discrete-Choice Analysis

24.3.2.1 The Effects of Policies

Consider a counterfactual policy p' such that the corresponding allocation of treatments $s_{p'}$ satisfies sequential randomization, as in (M-1):

(M-2) *For all treatment sequences s and all t,*

$$s_{p'}(t) \perp\!\!\!\perp (Y(t,s),\ldots,Y(\bar{T},s)) \mid (Y_{p'}^{t-1}, s_{p'}^{t-1} = s^{t-1}, Z).$$

The treatment assignment rule $s_{p'}$ is equivalent to what Gill and Robins (2001) call a "randomized plan". The outcome distribution under such a rule cannot be constructed by integrating the distributions of $\{Y(s)\}$ with respect to the distribution of $s_{p'}$, because there may be feedback from intermediate outcomes into treatment assignment. Instead, under the assumptions of the previous subsection and a support condition, we can use a version of the g-computation formula for randomized plans given by Gill and Robins to compute the distribution of outcomes under the policy p':[27]

$$\begin{aligned}\Pr\left(Y_{p'} = y|Z\right) = \sum_{s\in\mathscr{S}} & \Pr\left(Y_{p_0}(1) = y(1) \mid S(1) = s(1), Z\right) \Pr\left(s_{p'}(1) = s(1) \mid Z\right) \\ & \times \prod_{t=2}^{\bar{T}} \Big[\Pr\left(Y_{p_0}(t) = y(t) \mid Y_{p_0}^{t-1} = y^{t-1}, S^t = s^t, Z\right) \\ & \times \Pr\left(s_{p'}(t) = s(t) \mid Y_{p'}^{t-1} = y^{t-1}, s_{p'}^{t-1} = s^{t-1}, Z\right)\Big] \end{aligned} \tag{24.4}$$

In the special case of static matching on Z, so that $s_{p'} \perp\!\!\!\perp U \mid Z$, this simplifies to integrating the distribution of $Y_{p_0} \mid (S = s, Z)$ over the distribution of $s_{p'}|Z$:[28]

---

[27] The corresponding formula for the case with general treatment distributions is

$$\begin{aligned}&\Pr\left(Y_{p'} \in A|Z\right) \\ &\quad = \int_{y\in A}\int_{s\in\mathscr{S}} \Pr\left(Y_{p_0}(\bar{T}) \in \mathrm{d}y(\bar{T}) \mid Y_{p_0}^{\bar{T}-1} = y^{\bar{T}-1}, S^{\bar{T}} = s^{\bar{T}}, Z\right) \\ &\qquad\qquad \times \Pr\left(s_{p'}(\bar{T}) \in \mathrm{d}s(\bar{T}) \mid Y_{p'}^{\bar{T}-1} = y^{\bar{T}-1}, s_{p'}^{\bar{T}-1} = s^{\bar{T}-1}, Z\right) \\ &\qquad\qquad\qquad\qquad \vdots \\ &\qquad\qquad \times \Pr\left(Y_{p_0}(2) \in \mathrm{d}y(2) \mid Y_{p_0}(1) = y(1), S(1) = s(1), Z\right) \\ &\qquad\qquad \times \Pr\left(s_{p'}(2) \in \mathrm{d}s(2) \mid Y_{p'}(1) = y(1), s_{p'}(1) = s(1), Z\right) \\ &\qquad\qquad \times \Pr\left(Y_{p_0}(1) \in \mathrm{d}y(1) \mid S(1) = s(1), Z\right) \Pr\left(s_{p'}(1) \in \mathrm{d}s(1) \mid Z\right) .\end{aligned}$$

The support condition on $s_{p'}$ requires that, for each t, the distribution of $s_{p'}(t) \mid (Y_{p'}^{t-1} = y^{t-1}, s_{p'}^{t-1} = s^{t-1}, Z = z)$ is absolutely continuous with respect to the distribution of $S(t) \mid (Y_{p_0}^{t-1} = y^{t-1}, S^{t-1} = s^{t-1}, Z = z)$ for almost all (y^{t-1}, s^{t-1}, z) from the distribution of $(Y_{p_0}^{t-1}, S^{t-1}, Z)$.

[28] In the general case, this condition becomes

$$\Pr\left(Y_{p'} \in A|Z\right) = \int_{s\in\mathscr{S}} \Pr\left(Y_{p_0} \in A \mid S = s, Z\right) \Pr\left(s_{p'} \in \mathrm{d}s \mid Z\right) .$$

$$\Pr\left(Y_{p'} = y|Z\right) = \sum_{s \in \mathscr{S}} \Pr\left(Y_{p_0} = y \mid S = s, Z\right) \Pr\left(s_{p'} = s \mid Z\right) .$$

24.3.2.2 Policy Choice and Optimal Policies

We now consider the problem of choosing a policy p that is optimal according to some criterion. This problem is both of normative interest and of descriptive interest if actual policies are chosen to be optimal. We could, for example, study the optimal assignment a' of constraints and incentives to agents. Alternatively, we could assume that agents pick τ to maximize their utilities, and use the methods discussed in this section to model τ.

Under the policy invariance assumptions that underlie the treatment-effect approach, p only affects outcomes through its implied treatment allocation s_p. Thus, the problem of choosing an optimal policy boils down to choosing an optimal treatment allocation s_p under informational and other constraints specific to the problem at hand. For example, suppose that the planner and the agents have the same information, $\mathscr{I}_P(p) = \mathscr{I}_A(p)$, the planner assigns eligibility to a program by a, and agents fully comply, so that $\mathscr{B} = \mathscr{S}$ and $s_p = a$. Then, s_p can be any rule from $\mathscr{A}$ and is adapted to $\mathscr{I}_P(p) = \mathscr{I}_A(p)$.

For expositional convenience, we consider the optimal choice of a treatment assignment s_p adapted to the agent's information $\mathscr{I}_A(p)$ constructed earlier. We will use the word "agents" to refer to the decision maker in this problem, even though it can also apply to the planner's decision problem. An econometric approach to this problem is to estimate explicit dynamic choice models with explicit choice-outcome relationships. One emphasis in the literature is on Markovian discrete-choice models that satisfy Rust's (1987) conditional-independence assumption (see Rust, 1994). Other assumptions are made in the literature and we exposit them in Sect. 24.5.

Here, we explore the use of Rust's (1987) model as a model of treatment choice in a dynamic treatment-effect setting. In particular, we make explicit the additional structure that Rust's model, and in particular his conditional-independence assumption, imposes on Robins' dynamic potential-outcomes model. We follow Rust (1987) and focus on a finite treatment (control) space $\mathscr{S}$. In the notation of our model, payoffs are determined by the outcomes Y_p, treatment choices s_p, the "cost shocks" V, and the covariates Z. Rust (1987) assumes that $\{Y_p(t-1), V_t, Z\}$ is a controlled first-order Markov process, with initial condition $Y_p(0) \equiv 0$ and control s_p.[29] As before, V_t and Z are not causally affected by choices, but $Y_p(t)$ may causally depend on current and past choices. The agents choose a treatment assignment rule s_p that maximizes

[29] Rust (1987) assumes an infinite-horizon, stationary environment. Here, we present a finite-horizon version to facilitate a comparison with the dynamic potential-outcomes model and to link up with the analysis in Sect. 24.5.

$$E\left[\sum_{t=1}^{\bar{T}} \Upsilon_t\{Y_p(t-1), V_t, s_p(t), Z\} + \Upsilon_{\bar{T}+1}\{Y_p(\bar{T}), Z\} \,\middle|\, \mathscr{I}_A(1)\right], \tag{24.5}$$

for some (net and discounted) utility functions Υ_t and $\mathscr{I}_A(1) = \mathscr{I}_A(1,p)$, which is independent of p. $\Upsilon_{\bar{T}+1}\{Y_p(\bar{T}), Z\}$ is the terminal value. Under standard regularity conditions on the utility functions, we can solve backward for the optimal policy s_p. Because of Rust's Markov assumption, s_p has a Markovian structure,

$$s_p(t) \perp\!\!\!\perp (Y_p^{t-2}, V^{t-1}) \mid [Y_p(t-1), V_t, Z],$$

for $t = 2, \ldots, \bar{T}$, and $\{Y_p(t-1), V_t, Z\}$ is a first-order Markov process. Note that Z enters the model as an observed (by the econometrician) factor that shifts net utility. A key assumption embodied in the specification of (24.5) is time-separability of utility. Rust (1987), in addition, imposes separability between observed and unobserved state variables. This assumption plays no essential role in expositing the core ideas in Rust, and we will not make it here.

Rust's (1987) conditional-independence assumption imposes two key restrictions on the decision problem. It is instructive to consider these restrictions in isolation from Rust's Markov restriction. We make the model's causal structure explicit using the potential-outcomes notation. Note that the model has a recursive causal structure—the payoff-relevant state is controlled by current and past choices only—and satisfies no-anticipation condition (NA). Setting $Y(0,s) \equiv 0$ for specificity, and ignoring the Markov restriction, Rust's conditional-independence assumption requires, in addition to the assumption that there are no direct causal effects of choices on V, that

$$Y(s,t) \perp\!\!\!\perp V^t \mid \left[Y^{t-1}(s), Z\right], \text{ and} \tag{24.6}$$

$$V_{t+1} \perp\!\!\!\perp V^t \mid \left[Y^t(s), Z\right], \tag{24.7}$$

for all s and t. As noted by Rust (1987, p. 1011), condition (24.6) ensures that the observed (by the econometrician) controlled state evolves independently of the unobserved payoff-relevant variables. It is equivalent to (Florens and Mouchart, 1982)[30]

(M-3) $[Y(s,t), \ldots, Y(s,\bar{T})] \perp\!\!\!\perp V^t \mid \left[Y^{t-1}(s), Z\right]$ *for all t and s.*

In turn, (M-3) implies (M-1) and is equivalent to the assumption that (M-2) holds for all $s_{p'}$.[31]

Condition (24.7) excludes serial dependence of the unobserved payoff-relevant variables conditional on past outcomes. In contrast, Robins' g-computation framework allows for such serial dependence, provided that sequential randomization holds if serial dependence is present. For example, if $V \perp\!\!\!\perp U | Z$, then (M-1) and its variants hold without further assumptions on the time series structure of V_t.

[30] Note that (24.6) is a Granger (1969) noncausality condition stating that, for all s and conditional on Z, V does not cause $Y(s)$.

[31] If V has redundant components, that is components that do not nontrivially enter any assignment rule s_p, (M-3) imposes more structure, but structure that is irrelevant to the decision problem and its empirical analysis.

The first-order Markov assumption imposes additional restrictions on potential outcomes. These restrictions are twofold. First, potential outcomes follow a first-order Markov process. Second, $s(t)$ only directly affects the Markov transition from $Y(t,s)$ to $Y(t+1,s)$. This strengthens the no-anticipation assumption presented in Sect. 24.3.1.1. The Markov assumption also requires that V_{t+1} only depends on $Y(s,t)$, and not on $Y^{t-1}(s)$, given $Y(s,t)$.

In applications, we may assume that actual treatment assignment S solves the Markovian decision problem. Together with specifications of Υ_t, this further restricts the dynamic choice-outcome model. Alternatively, one could make other assumptions on S and use (24.5) to define and find an optimal, and typically counterfactual, assignment rule $s_{p'}$.

Our analysis shows that the substantial econometric literature on the structural empirical analysis of Markovian decision problems under conditional independence can be applied to policy evaluation under sequential randomization. Conversely, methods developed for potential-outcomes models with sequential randomization can be applied to learn about aspects of dynamic discrete-choice models. Murphy (2003) develops methods to estimate an optimal treatment assignment rule using Robins' dynamic potential-outcomes model with sequential randomization (M-2).

24.3.3 The Information Structure of Policies

One concern about methods for policy evaluation based on the potential-outcomes model is that potential outcomes are sometimes reduced-form representations of dynamic models of agent's choices. A policy maker choosing optimal policies typically faces a population of agents who act on the available information, and their actions in turn affect potential outcomes. For example, in terms of the model of Sect. 24.3.2.2, a policy may change financial incentives—the $b \in \mathscr{B}$ assigned through a could enter the net utilities Υ_t—and leave it to the agents to control outcomes by choosing treatment. In econometric policy evaluation, it is therefore important to carefully model the information $\mathscr{I}_A$ that accumulates to the agents in different program states and under different policies, separately from the policy maker's information $\mathscr{I}_P$.

This can be contrasted with common practice in biostatistics. Statistical analyses of the effects of drugs on health are usually concerned with the physician's (planner's) information and decision problem. Gill and Robins' (2001) sequential randomization assumption, for example, is often justified by the assumption that physicians base their treatment decisions on observable (by the analyst) information only. This literature, however, often ignores the possibility that many variables known to the physician may not be known to the observing statistician and that the agents being given drugs alter the protocols.

Potential outcomes will often depend on the agent's information. Failure to correctly model the information will often lead to violation of (NA) and failure of invariance. Potential outcomes may therefore not be valid inputs in a policy evaluation

study. A naive specification of potential outcomes would only index treatments by actual participation in, e.g., job search assistance or training programs. Such a naive specification is incomplete in the context of economies inhabited by forward-looking agents who make choices that affect outcomes. In specifying potential outcomes, we should not only consider the effects of actual program participation, but also the effects of the information available to agents about the program and policy. We now illustrate this point.

Example 24.2. Black, Smith, Berger, and Noel (2003) analyze the effect of compulsory training and employment services provided to unemployment insurance (UI) claimants in Kentucky on the exit rate from UI and earnings. In the program they study, letters are sent out to notify agents some time ahead whether they are selected to participate in the program. This information is recorded in a database and available to them. They can analyze the letter as part of a program that consists of information provision and subsequent participation in training. The main empirical finding of their paper is that the threat of future mandatory training conveyed by the letters is more effective in increasing the UI exit rate than training itself.

The data used by Black et al. (2003) are atypical of many economic data sets, because the data collectors carefully record the information provided to agents. This allows Black et al. to analyze the effects of the provision of information along with the effects of actual program participation. In many econometric applications, the information on the program under study is less rich. Data sets may provide information on actual participation in training programs and some background information on how the program is administered. Typically, however, the data do not record all of the letters sent to agents and do not record every phone conversation between administrators and agents. Then, the econometrician needs to make assumptions on how this information accumulates for agents. In many applications, knowledge of specific institutional mechanisms of assignment can be used to justify specific informational assumptions.

Example 24.3. Abbring, Van den Berg, and Van Ours (2005) analyze the effect of punitive benefits reductions, or sanctions, in Dutch UI on re-employment rates. In the Netherlands, UI claimants have to comply with certain rules concerning search behavior and registration. If a claimant violates these rules, a sanction may be applied. A sanction is a punitive reduction in benefits for some period of time and may be accompanied by increased levels of monitoring by the UI agency.[32] Abbring et al. (2005) use administrative data and know the re-employment duration, the duration at which a sanction is imposed if a sanction is imposed, and some background characteristics for each UI case.

Without prior knowledge of the Dutch UI system, an analyst might make a variety of informational assumptions. One extreme is that UI claimants know at the start of their UI spells that their benefits will be reduced at some specific duration if they are still claiming UI at that duration. This results in a UI system with entitlement periods that are tailored to individual claimants and that are set and revealed

[32] See Grubb (2000) for a review of sanction systems in the OECD.

at the start of the UI spells. In this case, claimants will change their labor market behavior from the start of their UI spell in response to the future benefits reduction (e.g., Mortensen, 1977). At another extreme, claimants receive no prior signals of impending sanctions and there are no anticipatory effects of actual benefits reductions. However, agents may still be aware of the properties of the sanctions process and to some extent this will affect their behavior. Abbring et al. (2005) analyze a search model with these features. Abbring and Van den Berg (2003b) provide a structural example where the data cannot distinguish between these two informational assumptions. Abbring et al. (2005) use institutional background information to argue in favor of the second informational assumption as the one that characterizes their data.

If data on information provision are not available and simplifying assumptions on the program's information structure cannot be justified, the analyst needs to model the information that accumulates to agents as an unobserved determinant of outcomes. This is the approach followed, and further discussed, in Sect. 24.5.

The information determining outcomes typically includes aspects of the policy. In Example 24.2, the letter announcing future training will be interpreted differently in different policy environments. If agents are forward looking, the letter will be more informative under a policy that specifies a strong relation between the letter and mandatory training in the population than under a policy that allocates letters and training independently. In Example 24.3, the policy is a monitoring regime. Potential outcomes are UI durations under different sanction times. A change in monitoring policy changes the value of unemployment. In a job-search model with forward-looking agents, agents will respond by changing their search effort and reservation wages, and UI duration outcomes will change. In either example, potential outcomes are not invariant to variation in the policy. In the terminology of Hurwicz (1962), the policy is not "structural" with regard to potential outcomes and violates the invariance assumptions presented in Heckman and Vytlacil (2007a). One must control for the effects of agents' information.

24.3.4 Selection on Unobservables

In econometric program evaluations, (sequentially) randomized assignment is unlikely to hold. We illustrate this in the models developed in Sect. 24.5. Observational data are characterized by a lot of heterogeneity among agents, as documented by the empirical examples in Abbring and Heckman (2007) and in Heckman et al. (1999). This heterogeneity is unlikely to be fully captured by the observed variables in most data sets. In a dynamic context, such unmeasured heterogeneity leads to violations of the assumptions of Gill and Robins (2001) and Rust (1987) that choices represent a sequential randomization. This is true even if the unmeasured variables only affect the availability of slots in programs but not outcomes directly. If agents are rational, forward-looking and observe at least some of the unmeasured variables that the econometrician does not, they will typically respond to these variables through their

choice of treatment and investment behavior. In this case, the sequential randomization condition fails.

For the same reason, identification based on instrumental variables is relatively hard to justify in dynamic models (Hansen and Sargent, 1980; Rosenzweig and Wolpin, 2000; Abbring and Van den Berg, 2005). If the candidate instruments only vary across persons but not over time for the same person, then they are not likely to be valid instruments because they affect expectations and future choices and may affect current potential outcomes. Instead of using instrumental variables that vary only across persons, we require instruments based on unanticipated person-specific shocks that affect treatment choices but not outcomes at each point in time. In the context of continuously assigned treatments, the implied data requirements seem onerous. To achieve identification, Abbring and Van den Berg (2003b) focus on regressor variation rather than exclusion restrictions in a sufficiently smooth model of continuous-time treatment effects. We discuss their analysis in Sect. 24.4. Heckman and Navarro (2007) show that curvature conditions, not exclusion restrictions, that result in the same variables having different effects on choices and outcomes in different periods, are motivated by economic theory and can be exploited to identify dynamic treatment effects in discrete time without literally excluding any variables. We discuss their analysis in Sect. 24.5. We now consider a formulation of the analysis in continuous time.

24.4 The Event-History Approach to Policy Analysis

The discrete-time models just discussed in Sect. 24.3 have an obvious limitation. Time is continuous and many events are best described by a continuous-time model. There is a rich field of continuous-time event-history analysis that has been adapted to conduct policy evaluation analysis.[33] For example, the effects of training and counseling on unemployment durations and job stability have been analyzed by applying event-history methods to data on individual labor-market and training histories (Ridder, 1986; Card and Sullivan, 1988; Gritz, 1993; Ham and LaLonde, 1996; Eberwein, Ham, and LaLonde, 1997; Bonnal, Fougère, and Sérandon, 1997). Similarly, the moral hazard effects of unemployment insurance have been studied by analyzing the effects of time-varying benefits on labor-market transitions (e.g., Meyer, 1990; Abbring et al., 2005; Van den Berg, Van der Klaauw, and Van Ours, 2004). In fields like epidemiology, the use of event-history models to analyze treatment effects is widespread (see, e.g., Andersen, Borgan, Gill, and Keiding, 1993; Keiding, 1999).

The event-history approach to program evaluation is firmly rooted in the econometric literature on state dependence (lagged dependent variables) and heterogeneity (Heckman and Borjas, 1980; and Heckman, 1981a). Event-history models along

[33] Abbring and Van den Berg (2004) discuss the relation between the event-history approach to program evaluation and more standard latent-variable and panel-data methods, with a focus on identification issues.

the lines of Heckman and Singer (1984, 1986) are used to jointly model transitions into programs and transitions into outcome states. Causal effects of programs are modelled as the dependence of individual transition rates on the individual history of program participation. Dynamic selection effects are modelled by allowing for dependent unobserved heterogeneity in both the program and outcome transition rates.

Without restrictions on the class of models considered, true state dependence and dynamic selection effects cannot be distinguished.[34] Any history dependence of current transition rates can be explained both as true state dependence and as the result of unobserved heterogeneity that simultaneously affects the history and current transitions. This is a dynamic manifestation of the problem of drawing causal inference from observational data. In applied work, researchers avoid this problem by imposing additional structure. A typical, simple, example is a mixed semi-Markov model in which the causal effects are restricted to program participation in the previous spell (e.g., Bonnal et al. 1997). There is a substantial literature on the identifiability of state-dependence effects and heterogeneity in duration and event-history models that exploit such additional structure (see Heckman and Taber, 1994; and Van den Berg, 2001 for reviews). Here, we provide discussion of some canonical cases.

24.4.1 Treatment Effects in Duration Models

24.4.1.1 Dynamically Assigned Binary Treatments and Duration Outcomes

We first consider the simplest case of mutual dependence of events in continuous time, involving only two binary events. This case is sufficiently rich to capture the effect of a dynamically assigned binary treatment on a duration outcome. Binary events in continuous time can be fully characterized by the time at which they occur and a structural model for their joint determination is a simultaneous-equations model for durations. We develop such a model along the lines of Abbring and Van den Berg (2003b). This model is an extension, with general marginal distributions and general causal and spurious dependence of the durations, of Freund's (1961) bivariate exponential model.

Consider two continuously-distributed random durations Y and S. We refer to one of the durations, S, as the time to treatment and to the other duration, Y, as the outcome duration. Such an asymmetry arises naturally in many applications. For example, in Abbring et al.'s (2005) study of unemployment insurance, the treatment is a punitive benefits reduction (sanction) and the outcome re-employment. The re-employment process continues after imposition of a sanction, but the sanctions process is terminated by re-employment. The current exposition, however, is symmetric and unifies both cases. It applies to both the asymmetric setup of the sanctions example and to applications in which both events may causally affect the other event.

[34] See Heckman and Singer (1986).

Let $Y(s)$ be the potential outcome duration that would prevail if the treatment time is externally set to s. Similarly, let $S(y)$ be the potential treatment time resulting from setting the outcome duration to y. We assume that *ex ante* heterogeneity across agents is fully captured by observed covariates X and unobserved covariates V, assumed to be external and temporally invariant. Treatment causally affects the outcome duration through its hazard rate. We denote the hazard rate of $Y(s)$ at time t for an agent with characteristics (X,V) by $\theta_Y(t|s,X,V)$. Similarly, outcomes affect the treatment times through its hazard $\theta_S(t|y,X,V)$. Causal effects on hazard rates are produced by recursive economic models driven by point processes, such as search models. We provide an example below, and further discussion in Sect. 24.4.3.

Without loss of generality, we partition V into (V_S,V_Y) and assume that $\theta_Y(t|s,X,V)=\theta_Y(t|s,X,V_Y)$ and $\theta_S(t|y,X,V)=\theta_S(t|y,X,V_S)$. Intuitively, V_S and V_Y are the unobservables affecting, respectively, treatment and outcome, and the joint distribution of (V_S,V_Y) is unrestricted. In particular, V_S and V_Y may have elements in common.

The corresponding integrated hazard rates are defined by $\Theta_Y(t \mid s,X,V_Y)=\int_0^t \theta_Y(u \mid s,X,V_Y)\mathrm{d}u$ and $\Theta_S(t \mid y,X,V_S)=\int_0^t \theta_S(u \mid y,X,V_S)\mathrm{d}u$. For expositional convenience, we assume that these integrated hazards are strictly increasing in t. We also assume that they diverge to ∞ as $t\to\infty$, so that the duration distributions are non-defective.[35] Then, $\Theta_Y(Y(s) \mid s,X,V_Y)$ and $\Theta_S(S(y) \mid y,X,V_S)$ are unit exponential for all $y,s\in\mathbb{R}_+$.[36] This implies the following model of potential outcomes and treatments,[37]

$$Y(s)=y(s,X,V_Y,\varepsilon_Y) \text{ and } S(y)=s(y,X,V_S,\varepsilon_S)\,,$$

for some unit exponential random variables ε_Y and ε_S that are independent of (X,V), $y=\Theta_Y^{-1}$, and $s=\Theta_S^{-1}$.

The exponential errors ε_Y and ε_S embody the *ex post* shocks that are inherent to the individual hazard processes, that is the randomness in the transition process after conditioning on covariates X and V and survival. We assume that $\varepsilon_Y \perp\!\!\!\perp \varepsilon_S$, so that $\{Y(s)\}$ and $\{S(y)\}$ are only dependent through the observed and unobserved covariates (X,V). This conditional-independence assumption is weaker than the conditional-independence assumption underlying the analysis of Sect. 24.3 and used

[35] Abbring and Van den Berg (2003b) allow for defective distributions, which often have structural interpretations. For example, some women never have children and some workers will never leave a job. See Abbring (2002) for discussion.

[36] Let $T \mid X$ be distributed with density $f(t|X)$, non-defective cumulative distribution function $F(t|X)$, and hazard rate $\theta(t|X)=f(t|X)/[1-F(t|X)]$. Then, $\int_0^T \theta(t|X)\mathrm{d}t=-\ln[1-F(T|X)]$ is a unit exponential random variable that is independent of X.

[37] The causal hazard model only implies that the distributions of ε_Y and ε_S are invariant across assigned treatments and outcomes, respectively; their realizations may not be. This is sufficient for the variation of $y(s,X,V_Y,\varepsilon_Y)$ with s and of $s(y,X,V_S,\varepsilon_S)$ with y to have a causal interpretation. The further restriction that the random variables ε_Y and ε_S are invariant is made for simplicity, and is empirically innocuous. See Abbring and Van den Berg (2003b) for details and Freedman (2004) for discussion.

in matching, because it allows for conditioning on the invariant unobservables V. It shares this feature with the discrete-time models developed in Sect. 24.5.

We assume a version of the no-anticipation condition of Sect. 24.3.1.1: for all $t \in \mathbb{R}_+$,

$$\theta_Y(t|s,X,V_Y) = \theta_Y(t|s',X,V_Y) \text{ and } \theta_S(t|y,X,V_S) = \theta_S(t|y',X,V_S)$$

for all $s,s',y,y' \in [t,\infty)$. This excludes effects of anticipation of the treatment on the outcome. Similarly, there can be no anticipation effects of future outcomes on the treatment time hazard.

Example 24.4. Consider a standard search model describing the job search behavior of an unemployed individual (e.g., Mortensen, 1986) with characteristics (X,V). Job offers arrive at a rate $\lambda > 0$ and are random draws from a given distribution F. Both λ and F may depend on (X,V), but, for notational simplicity we suppress all explicit representations of conditioning on (X,V) throughout this example. An offer is either accepted or rejected. A rejected offer cannot be recalled at a later time. The individual initially receives a constant flow of unemployment-insurance benefits. However, the individual faces the risk of a sanction—a permanent reduction of his benefits to some lower, constant level—at some point during his unemployment spell. During the unemployment spell, sanctions arrive independently of the job-offer process at a constant rate $\mu > 0$. The individual cannot foresee the exact time a sanction is imposed, but he knows the distribution of these times.[38] The individual chooses a job-acceptance rule so as to maximize his expected discounted lifetime income. Under standard conditions, this is a reservation-wage rule: at time t, the individual accepts each wage of $\underline{w}(t)$ or higher. The corresponding re-employment hazard rate is $\lambda(1-F(\underline{w}(t)))$. Apart from the sanction, which is not foreseen and arrives at a constant rate during the unemployment spell, the model is stationary. This implies that the reservation wage is constant, say equal to $\underline{w}_0$, up to and including time s, jumps to some lower level $\underline{w}_1 < \underline{w}_0$ at time s and stays constant at $\underline{w}_1$ for the remainder of the unemployment spell if benefits would be reduced at time s.

The model is a version of the simultaneous-equations model for durations. To see this, let Y be the re-employment duration and S the sanction time. The potential-outcome hazards are

$$\theta_Y(t|s) = \begin{cases} \lambda_0 \text{ if } 0 \le t \le s \\ \lambda_1 \text{ if } t > s\,, \end{cases}$$

where $\lambda_0 = \lambda\,[1-F(\underline{w}_0)]$ and $\lambda_1 = \lambda\,[1-F(\underline{w}_1)]$, and clearly $\lambda_1 \ge \lambda_0$. Similarly, the potential-treatment time hazards are $\theta_S(t|y) = \mu$ if $0 \le t \le y$, and 0 otherwise. Note that the no-anticipation condition follows naturally from the recursive structure of the economic decision problem in this case, in which we have properly accounted for all relevant components of agent information sets. Furthermore, the assumed

[38] This is a rudimentary version of the search model with punitive benefits reductions, or sanctions, of Abbring et al. (2005). The main difference is that in the present version of the model the sanctions process cannot be controlled by the agent.

independence of the job offer and sanction processes at the individual level for given (X,V) implies that $\varepsilon_Y \perp\!\!\!\perp \varepsilon_S$.

The actual outcome and treatment are related to the potential outcomes and treatments by $S = S(Y)$ and $Y = Y(S)$. The no-anticipation assumption ensures that this system has a unique solution (Y,S) by imposing a recursive structure on the underlying transition processes. Without anticipation effects, current treatment and outcome hazards only depend on past outcome and treatment events, and the transition processes evolve recursively (Abbring and Van den Berg, 2003b). Together with a distribution $G(\cdot \mid X)$ of $V \mid X$, this gives a non-parametric structural model of the distribution of $(Y,S) \mid X$ that embodies general simultaneous causal dependence of Y and S, dependence of (Y,X) on observed covariates X, and general dependence of the unobserved errors V_Y and V_S.

There are two reasons for imposing further restrictions on this model. First, it is not identified from data on (Y,S,X). Take a version of the model with selection on unobservables ($V_Y \not\perp\!\!\!\perp V_S \mid X$) and consider the distribution of $(Y,S)|X$ generated by this version of the model. Then, there exists an alternative version of the model that satisfies both no-anticipation and $V_Y \perp\!\!\!\perp V_S \mid X$, and that generates the same distribution of $(Y,S)|X$ (Abbring and Van den Berg, 2003b, Proposition 1). In other words, for each version of the model with selection on unobservables and anticipation effects, there is an observationally-equivalent model version that satisfies no-anticipation and conditional randomization. This is a version of the nonidentification result discussed in Sect. 24.3.1.

Second, even if we ensure nonparametric identification by assuming no-anticipation and conditional randomization, we cannot learn about the agent-level causal effects embodied in y and s without imposing even further restrictions. At best, under regularity conditions we can identify $\theta_Y(t|s,X) = E[\theta_Y(t|s,X,V_Y)|X,Y(s) \geq t]$ and $\theta_S(t|y,X) = E[\theta_S(t|y,X,V_S)|X,S(y) \geq t]$ from standard hazard regressions (e.g., Andersen et al., 1993; Fleming and Harrington, 1991). Thus, we can identify the distributions of $Y(s)|X$ and $S(y)|X$, and therefore solve the selection problem if we are only interested in these distributions. However, if we are also interested in the causal effects on the corresponding hazard rates for given X,V, we face an additional dynamic selection problem. The hazards of the identified distributions of $Y(s)|X$ and $S(y)|X$ only condition on observed covariates X, and not on unobserved covariates V, and are confounded with dynamic selection effects (Heckman and Borjas, 1980; Heckman and Singer, 1986; Meyer, 1996; Abbring and Van den Berg, 2005). For example, the difference between $\theta_Y(t|s,X)$ and $\theta_Y(t|s',X)$ does not only reflect agent-level differences between $\theta_Y(t|s,X,V_Y)$ and $\theta_Y(t|s',X,V_Y)$, but also differences in the subpopulations of survivors $\{X,Y(s) \geq t\}$ and $\{X,Y(s') \geq t\}$ on which the hazards are computed.

In the next two subsections, we discuss what can be learned about treatment effects in duration models under additional model restrictions. We take the no-anticipation assumption as fundamental. As explained in Sect. 24.3, this requires that we measure and include in our model all relevant information needed to define potential outcomes. However, we relax the randomization assumption. We first consider Abbring and Van den Berg's (2003b) analysis of identification without

exclusion restrictions. They argue that these results are useful, because exclusion restrictions are hard to justify in an inherently dynamic setting with forward-looking agents. Abbring and Van den Berg (2005) further clarify this issue by studying inference for treatment effects in duration models using a social experiment. We discuss what can be learned from such experiments at the end of this section.

24.4.1.2 Identifiability Without Exclusion Restrictions

Abbring and Van den Berg consider an extension of the multivariate Mixed Proportional Hazard (MPH) model (Lancaster, 1979) in which the hazard rates of $Y(s) \mid (X,V)$ and $S(y) \mid (X,V)$ are given by

$$\theta_Y(t \mid s,X,V) = \begin{cases} \lambda_Y(t)\phi_Y(X)V_Y & \text{if } t \leq s \\ \lambda_Y(t)\phi_Y(X)\delta_Y(t,s,X)V_Y & \text{if } t > s \end{cases} \tag{24.8}$$

and

$$\theta_S(t \mid y,X,V) = \begin{cases} \lambda_S(t)\phi_S(X)V_S & \text{if } t \leq y \\ \lambda_S(t)\phi_S(X)\delta_S(t,y,X)V_S & \text{if } t > y, \end{cases} \tag{24.9}$$

respectively, and $V = (V_S,V_Y)$ is distributed independently of X. The baseline hazards $\lambda_Y : \mathbb{R}_+ \to (0,\infty)$ and $\lambda_S : \mathbb{R}_+ \to (0,\infty)$ capture duration dependence of the individual transition rates. The integrated hazards are $\Lambda_Y(t) = \int_0^t \lambda_Y(\tau)\mathrm{d}\tau < \infty$ and $\Lambda_S(t) = \int_0^t \lambda_S(\tau)\mathrm{d}\tau < \infty$, for all $t \in \mathbb{R}_+$. The regressor functions $\phi_Y : \mathscr{X} \to (0,\infty)$ and $\phi_S : \mathscr{X} \to (0,\infty)$ are assumed to be continuous, with $\mathscr{X} \subset \mathbb{R}^q$ the support of X. In empirical work, these functions are frequently specified as $\phi_Y(x) = \exp(x'\beta_Y)$ and $\phi_S(x) = \exp(x'\beta_S)$ for some parameter vectors β_Y and β_S. We will not make such parametric assumptions. Note that the fact that both regressor functions are defined on the same domain $\mathscr{X}$ is not restrictive, because each function ϕ_Y and ϕ_S can "select" certain elements of X by being trivial functions of the other elements. In the parametric example, the vector β_Y would only have nonzero elements for those regressors that matter to the outcome hazard. The functions δ_Y and δ_S capture the causal effects. Note that $\delta_Y(t,s,X)$ only enters $\theta_Y(t \mid s,X,V)$ at durations $t > s$, so that the model satisfies no anticipation of treatment assumption (NA). Similarly, it satisfies no anticipation of outcomes and has a recursive causal structure as required by the no-anticipation assumption. If $\delta_Y = 1$, treatment is ineffective; if δ_Y is larger than 1, it stochastically reduces the remaining outcome duration.

Note that this model allows δ_Y and δ_S to depend on elapsed duration t, past endogenous events, and the observed covariates X, but not on V. Abbring and Van den Berg also consider an alternative model that allows δ_Y and δ_S to depend on unobservables in a general way, but not on past endogenous events.

Abbring and Van den Berg show that these models are nonparametrically identified from single-spell data under the conditions for the identification of competing-risks models based on the multivariate MPH model given by Abbring and Van den Berg (2003a). Among other conditions are the requirements that there is some

independent local variation of the regressor effects in both hazard rates and a finite-mean restriction on V, which are standard in the analysis of multivariate MPH models. With multiple-spell data, most of these assumptions, and the MPH structure, can be relaxed (Abbring and Van den Berg, 2003b).

The models can be parameterized in a flexible way and estimated by maximum likelihood. Typical parameterizations involve linear-index structures for the regressor and causal effects, a discrete distribution G, and piecewise-constant baseline hazards λ_S and λ_Y. Abbring and Van den Berg (2003c) develop a simple graphical method for inference on the sign of $\ln(\delta_Y)$ in the absence of regressors. Abbring et al. (2005) present an empirical application.

24.4.1.3 Inference Based on Instrumental Variables

The concerns expressed in Sect. 24.3.4 about the validity of exclusion restrictions in dynamic settings carry over to event-history models.

Example 24.5. A good illustration of this point is offered by the analysis of Eberwein et al. (1997), who study the effects of a training program on labor-market transitions. Their data are particularly nice, as potential participants are randomized into treatment and control groups at some baseline point in time. This allows them to estimate the effect of intention to treat (with training) on subsequent labor-market transitions. This is directly relevant to policy evaluation in the case that the policy involves changing training enrollment through offers of treatment which may or may not be accepted by agents.

However, Eberwein et al. are also interested in the effect of actual participation in the training program on post program labor-market transitions. This is a distinct problem, because compliance with the intention-to-treat protocol is imperfect. Some agents in the control group are able to enroll in substitute programs, and some agents in the treatment group choose never to enroll in a program at all. Moreover, actual enrollment does not take place at the baseline time, but is dispersed over time. Those in the treatment group are more likely to enroll earlier. This fact, coupled with the initial randomization, suggests that the intention-to-treat indicator might be used as an instrument for identifying the effect of program participation on employment and unemployment spells.

The dynamic nature of enrollment into the training program, and the event-history focus of the analysis complicate matters considerably. Standard instrumental-variables methods cannot be directly applied. Instead, Eberwein et al. use a parametric duration model for pre and post program outcomes that excludes the intention-to-treat indicator from directly determining outcomes. They specify a duration model for training enrollment that includes an intention-to-treat indicator as an explanatory variable, and specify a model for labor-market transitions that excludes the intention-to-treat indicator and imposes a no-anticipation condition on the effect of actual training participation on labor-market transitions. Such a model is consistent with an environment in which agents cannot perfectly foresee the actual training time they will be assigned and in which they do not respond to information

about this time revealed by their assignment to an intention-to-treat group. This is a strong assumption. In a search model with forward-looking agents, for example, such information would typically affect the *ex ante* values of unemployment and employment. Then, it would affect the labor-market transitions before actual training enrollment through changes in search efforts and reservation wages, unless these are both assumed to be exogenous. An assumption of perfect foresight on the part of the agents being studied only complicates matters further.

Abbring and Van den Berg (2005) study what can be learned about dynamically assigned programs from social experiments if the intention-to-treat instrument cannot be excluded from the outcome equation. They develop bounds, tests for unobserved heterogeneity, and point-identification results that extend those discussed in this section.[39]

24.4.2 Treatment Effects in More General Event-History Models

It is instructive to place the causal duration models developed in Sect. 24.4.1 in the more general setting of event-history models with state dependence and heterogeneity. We do this following Abbring's (2008) analysis of the mixed semi-Markov model.

24.4.2.1 The Mixed Semi-Markov Event-History Model

The model is formulated in a fashion that is analogous to the frameworks of Heckman and Singer (1986). The point of departure is a continuous-time stochastic process assuming values in a finite set $\mathscr{S}$ at each point in time. We will interpret realizations of this process as agents' event histories of transitions between states in the state space $\mathscr{S}$.

Suppose that event histories start at real-valued random times T_0 in an $\mathscr{S}$-valued random state S_0, and that subsequent transitions occur at random times $T_1, T_2, \ldots$ such that $T_0 < T_1 < T_2 < \cdots$. Let S_l be the random destination state of the transition at T_l. Taking the sample paths of the event-history process to be right-continuous, we have that S_l is the state occupied in the interval $[T_l, T_{l+1})$.

Suppose that heterogeneity among agents is captured by vectors of time-constant observed covariates X and unobserved covariates V.[40] In this case, state dependence in the event-history process for given individual characteristics X, V has a causal

[39] In the special case that a static treatment, or treatment plan, is assigned at the start of the spell, standard instrumental-variables methods may be applied. See Abbring and Van den Berg (2005).

[40] We restrict attention to time-invariant observed covariates for expositional convenience. The analysis can easily be adapted to more general time-varying external covariates. Restricting attention to time-constant regressors is a worst-case scenario for identification. External time variation in observed covariates aids identification (Heckman and Taber, 1994).

interpretation.[41] We structure such state dependence by assuming that the event-history process conditional on X,V is a time-homogeneous semi-Markov process. Conditional on X,V the length of a spell in a state and the destination state of the transition ending that spell depend only on the past through the current state. In our notation, $(\Delta T_l, S_l) \perp\!\!\!\perp \{(T_i, S_i), i = 0, \ldots, l-1\} \mid S_{l-1}, X, V$, where $\Delta T_l = T_l - T_{l-1}$ is the length of spell l. Also, the distribution of $(\Delta T_l, S_l)|S_{l-1}, X, V$ does not depend on l. Note that, conditional on X,V, $\{S_l, l \geq 0\}$ is a time-homogeneous Markov chain under these assumptions.

Non-trivial dynamic selection effects arise because V is not observed. The event-history process conditional on observed covariates X only is a mixed semi-Markov process. If V affects the initial state S_0, or transitions from it, subpopulations of agents in different states at some time t typically have different distributions of the unobserved characteristics V. Therefore, a comparison of the subsequent transitions in two such subpopulations does not only reflect state dependence, but also sorting of agents with different unobserved characteristics into the different states they occupy at time t.

We model $\{(\Delta T_l, S_l), l \geq 1\}|T_0, S_0, X, V$ as a repeated competing-risks model. Due to the mixed semi-Markov assumption, the latent durations corresponding to transitions into the possible destination states in the lth spell only depend on the past through the current state S_{l-1}, conditional on X,V. This implies that we can fully specify the repeated competing-risks model by specifying a set of origin-destination-specific latent durations, with corresponding transition rates. Let T_{jk}^l denote the latent duration corresponding to the transition from state j to state k in spell l. We explicitly allow for the possibility that transitions between certain (ordered) pairs of states may be impossible. To this end, define the correspondence $\mathscr{Q} : \mathscr{S} \to \sigma(\mathscr{S})$ assigning to each $s \in \mathscr{S}$ the set of all destination states to which transitions are made from s with positive probability.[42] Here, $\sigma(\mathscr{S})$ is the set of all subsets of $\mathscr{S}$ (the "power set" of $\mathscr{S}$). Then, the length of spell l is given by $\Delta T_l = \min_{s \in \mathscr{Q}(S_{l-1})} T_{S_{l-1}s}^l$, and the destination state by $S_l = \arg\min_{s \in \mathscr{Q}(S_{l-1})} T_{S_{l-1}s}^l$.

We take the latent durations to be mutually independent, jointly independent of T_0, S_0, and identically distributed across spells l, all conditional on X,V. This reflects both the mixed semi-Markov assumption and the additional assumption that all dependence between the latent durations corresponding to the competing risks in a given spell l is captured by the observed regressors X and the unobservables V. This is a standard assumption in econometric duration analysis, which, with the semi-Markov assumption, allows us to characterize the distribution of $\{(\Delta T_l, S_l), l \geq 1\}|T_0, S_0, X, V$ by specifying origin-destination-specific hazards $\theta_{jk}(t|X,V)$ for the marginal distributions of $T_{jk}^l|X,V$.

[41] We could make this explicit by extending the potential-outcomes model of Sect. 24.4.1.2 to the general event-history setup. However, this would add a lot of complexity, but little extra insight.

[42] Throughout this section, we assume that $\mathscr{Q}$ is known. It is important to note, however, that $\mathscr{Q}$ can actually be identified trivially in all cases considered.

We assume that the hazards $\theta_{jk}(t|X,V)$ are of the mixed proportional hazard (MPH) type:[43]

$$\theta_{jk}(t|X,V) = \begin{cases} \lambda_{jk}(t)\phi_{jk}(X)V_{jk} & \text{if } k \in \mathscr{Q}(j) \\ 0 & \text{otherwise.} \end{cases} \tag{24.10}$$

The baseline hazards $\lambda_{jk} : \mathbb{R}_+ \to (0,\infty)$ have integrated hazards $\Lambda_{jk}(t) = \int_0^t \lambda_{jk}(\tau)\mathrm{d}\tau < \infty$, for all $t \in \mathbb{R}_+$. The regressor functions $\phi_{jk} : \mathscr{X} \to (0,\infty)$ are assumed to be continuous. Finally, the $(0,\infty)$-valued random variable V_{jk} is the scalar component of V that affects the transition from state j to state k. Note that we allow for general dependence between the components of V. This way, we can capture, for example, that agents with lower re-employment rates have higher training enrolment rates.

This model fully characterizes the distribution of the transitions $\{(\Delta T_l, S_l), l \geq 1\}$ conditional on the initial conditions T_0, S_0 and the agent's characteristics X,V. A complete model of the event histories $\{(T_l,S_l), l \geq 0\}$ conditional on X,V would in addition require a specification of the initial conditions T_0, S_0 for given X,V. It is important to stress here that T_0, S_0 are the initial conditions of the event-history process itself, and should not be confused with the initial conditions in a particular sample (which we will discuss below). In empirical work, interest in the dependence between start times T_0 and characteristics X,V is often limited to the observation that the distribution of agents' characteristics may vary over cohorts indexed by T_0. The choice of initial state S_0 may in general be of some interest, but is often trivial. For example, we could model labor-market histories from the calendar time T_0 at which agents turn 15 onwards. In an economy with perfect compliance to a mandatory schooling up to age 15, the initial state S_0 would be "(mandatory) schooling" for all. Therefore, we will not consider a model of the event history's initial conditions, but instead focus on the conditional model of subsequent transition histories.

Because of the semi-Markov assumption, the distribution of $\{(\Delta T_l, S_l), l \geq 1\}|$ T_0, S_0, X, V only depends on S_0, and not T_0. Thus, T_0 only affects observed event histories through cohort effects on the distribution of unobserved characteristics V. The initial state S_0, on the other hand, may both have causal effects on subsequent transitions and be informative on the distribution of V. For expositional clarity, we assume that $V \perp\!\!\!\perp (T_0, S_0, X)$. This is true, for example, if all agents start in the same state, so that S_0 is degenerate, and V is independent of the start date T_0 and the observed covariates X.

An econometric model for transition histories conditional on the observed covariates X can be derived from the model of $\{(\Delta T_l, S_l), l \geq 1\}|S_0, X, V$ by integrating out V. The exact way this should be done depends on the sampling scheme used. Here, we focus on sampling from the population of event-histories. We assume that we observe the covariates X, the initial state S_0, and the first $\bar{L}$ transitions from there. Then, we can model these transitions for given S_0, X by integrating the conditional model over the distribution of V.

[43] Proportionality can be relaxed if we have data on sufficiently long event-histories. See Honoré (1993) and Abbring and Van den Berg (2003a,b) for related arguments for various multi-spell duration models.

Abbring (2008) discusses more complex, and arguably more realistic, sampling schemes. For example, when studying labor-market histories we may randomly sample from the stock of the unemployed at a particular point in time. Because the unobserved component V affects the probability of being unemployed at the sampling date, the distribution of $V|X$ in the stock sample does not equal its population distribution. This is again a dynamic version of the selection problem. Moreover, in this case, we typically do not observe an agent's entire labor-market history from T_0 onwards. Instead, we may have data on the time spent in unemployment at the sampling date and on labor-market transitions for some period after the sampling date. This "initial-conditions problem" complicates matters further (Heckman, 1981b).

In the next two subsections, we first discuss some examples of applications of the model and then review a basic identification result for the simple sampling scheme above.

24.4.2.2 Applications to Program Evaluation

Several empirical papers study the effect of a single treatment on some outcome duration or set of transitions. Two approaches can be distinguished. In the first approach, the outcome and treatment processes are explicitly and separately specified. The second approach distinguishes treatment as one state within a single event-history model with state dependence.

The first approach is used in a variety of papers in labor economics. Eberwein et al. (1997) specify a model for labor market transitions in which the transition intensities between various labor market states (not including treatment) depend on whether someone has been assigned to a training program in the past or not. Abbring et al. (2005) and Van den Berg et al. (2004) specify a model for re-employment durations in which the re-employment hazard depends on whether a punitive benefits reduction has been imposed in the past. Similarly, Van den Berg, Holm, and Van Ours (2002) analyze the duration up to transition into medical trainee positions and the effect of an intermediate transition into a medical assistant position (a "stepping-stone job") on this duration. In all of these papers, the outcome model is complemented with a hazard model for treatment choice.

These models fit into the framework of Sect. 24.4.1.2 or a multi-state extension thereof. We can rephrase the class of models discussed in Sect. 24.4.1.2 in terms of a simple event-history model with state-dependence as follows. Distinguish three states, untreated (O), treated (P) and the exit state of interest (E), so that $\mathscr{S} = \{O,P,E\}$. All subjects start in O, so that $S_0 = O$. Obviously, we do not want to allow for all possible transitions between these three states. Instead, we restrict the correspondence $\mathscr{Q}$ representing the possible transitions as follows:

$$\mathscr{Q}(s) = \begin{cases} \{P,E\} & s = O, \\ \{E\} & \text{if } s = P, \\ \emptyset & s = E. \end{cases}$$

State dependence of the transition rates into E captures treatment effects in the sense of Sect. 24.4.1.2. Not all models in Abbring and Van den Berg (2003b) are included in the semi-Markov setup discussed here. In particular, in this paper we do not allow the transition rate from P to E to depend on the duration spent in O. This extension with "lagged duration dependence" (Heckman and Borjas, 1980) would be required to capture one variant of their model.

The model for transitions from "untreated" (O) is a competing risks model, with program enrolment (transition to P) and employment (E) competing to end the untreated spell. If the unobservable factor V_{OE} that determines transitions to employment and the unobservable factor V_{OP} affecting program enrolment are dependent, then program enrolment is selective in the sense that the initial distribution of V_{OE}—and also typically that of V_{PE}—among those who enroll at a given point in time does not equal its distribution among survivors in O up to that time.[44]

The second approach is used by Gritz (1993) and Bonnal et al. (1997), among others. Consider the following simplified setup. Suppose workers are either employed (E), unemployed (O), or engaged in a training program (P). We can now specify a transition process among these three labor market states in which a causal effect of training on unemployment and employment durations is modeled as dependence of the various transition rates on the past occurrence of a training program in the labor market history. Bonnal et al. (1997) only have limited information on agents' labor-market histories before the sample period. Partly to avoid difficult initial-conditions problems, they restrict attention to "first order lagged occurrence dependence" (Heckman and Borjas, 1980) by assuming that transition rates only depend on the current and previous states occupied. Such a model is not directly covered by the semi-Markov model, but with a simple augmentation of the state space it can be covered. In particular, we have to include lagged states in the state space on which the transition process is defined. Because there is no lagged state in the event-history's first spell, initial states should be defined separately. So, instead of just distinguishing states in $\mathscr{S}^* = \{E, O, P\}$, we distinguish augmented states in $\mathscr{S} = \{(s,s') \in (\mathscr{S}^* \cup \{I\}) \times \mathscr{S}^* : s \neq s'\}$. Then, (I,s), $s \in \mathscr{S}^*$, denote the initial states, and $(s,s') \in \mathscr{S}$ the augmented state of an agent who is currently in s' and came from $s \neq s'$. In order to preserve the interpretation of the model as a model of lagged occurrence dependence, we have to exclude certain transitions by specifying

$$\mathscr{Q}(s,s') = \{(s',s''), s'' \in \mathscr{S}^* \backslash \{s'\}\} .$$

This excludes transitions to augmented states that are labeled with a lagged state different from the origin state. Also, it ensures that agents never return to an initial state. For example, from the augmented state (O,P)—previously unemployed and currently enrolled in a program—only transitions to augmented states (P,s'')—previously enrolled in a program and currently in s''—are possible. Moreover, it is not possible to be currently employed and transiting to initially unemployed, (I,O).

[44] Note that, in addition, the survivors in O themselves are a selected subpopulation. Because V affects survival in O, the distribution of V among survivors in O is not equal to its population distribution.

Rather, an employed person who loses her job would transit to (E,O)—currently unemployed and previously employed.

The effects of training, for example, are now modeled as simple state-dependence effects. For example, the effect of training on the transition rate from unemployment to employment is simply the contrast between the individual transition rate from (E,O) to (O,E) and the transition rate from (P,O) to (O,E). Dynamic selection into the augmented states (E,O) and (P,O), as specified by the transition model, confounds the empirical analysis of these training effects. Note that due to the fact that we have restricted attention to first-order lagged occurrence dependence, there are no longer-run effects of training on transition rates from unemployment to employment.

24.4.2.3 Identification Without Exclusion Restrictions

In this section, we state a basic identification result for the following sampling scheme. Suppose that the economist randomly samples from the population of event-histories, and that we observe the first $\bar{L}$ transitions (including destinations) for each sampled event-history, with the possibility that $\bar{L} = \infty$.[45] Thus, we observe a random sample of $\{(T_l, S_l), l \in \{0,1,\ldots,\bar{L}\}\}$, and X.

First note that we can only identify the determinants of θ_{jk} for transitions (j,k) that occur with positive probability among the first $\bar{L}$ transitions. Moreover, without further restrictions, we can only identify the joint distribution of a vector of unobservables corresponding to (part of) a sequence of transitions that can be observed among the first $\bar{L}$ transitions.

With this qualification, identification can be proved by extending Abbring and Van den Berg's (2003a) analysis of the MPH competing risks model to the present setting. This analysis assumes that transition rates have an MPH functional form. Identification again requires specific moments of V to be finite, and independent local variation in the regressor effects.

24.4.3 A Structural Perspective

Without further restrictions, the causal duration model of Sect. 24.4.1.1 is versatile. It can be generated as the reduced form of a wide variety of continuous-time economic models driven by point processes. Leading examples are sequential job-search models in which job-offer arrival rates, and other model parameters, depend on agent characteristics (X,V) and policy interventions (see, e.g., Mortensen, 1986, and Example 24.4).

The MPH restriction on this model, however, is hard to justify from economic theory. In particular, nonstationary job-search models often imply interactions

[45] Note that this assumes away econometric initial-conditions problems of the type previously discussed.

between duration and covariate effects; the MPH model only results under strong assumptions (Heckman and Singer, 1986; Van den Berg, 2001). Similarly, an MPH structure is hard to generate from models in which agents learn about their individual value of the model's structural parameters, that is about (X,V), through Bayesian updating.

An alternative class of continuous-time models, not discussed in this chapter, specifies durations as the first time some Gaussian or more general process crosses a threshold. Such models are closely related to a variety of dynamic economic models. They have attracted recent attention in statistics (see, e.g., Aalen and Gjessing, 2004). Abbring (2007) analyzes identifiability of "mixed hitting-time models", continuous-time threshold-crossing models in which the parameters depend on observed and unobserved covariates, and discusses their link with optimizing models in economics. This is a relatively new area of research, and a full development is beyond the scope of this paper. It extends to a continuous-time framework the dynamic threshold crossing model developed in Heckman (1981a,b) that is used in the next subsection of this chapter.

We now discuss a complementary discrete-time approach where it is possible to make many important economic distinctions that are difficult to make in the setting of continuous-time models and to avoid some difficult measure-theoretic problems. In the structural version, it is possible to specify precisely agent information sets in a fashion that is not possible in conventional duration models.

24.5 Dynamic Discrete Choice and Dynamic Treatment Effects

Heckman and Navarro (2007) and Cunha, Heckman, and Navarro (2007) present econometric models for analyzing time to treatment and the consequences of the choice of a particular treatment time. Treatment may be a medical intervention, stopping schooling, opening a store, conducting an advertising campaign at a given date or renewing a patent. Associated with each treatment time, there can be multiple outcomes. They can include a vector of health status indicators and biomarkers; lifetime employment and earnings consequences of stopping at a particular grade of schooling; the sales revenue and profit generated from opening a store at a certain time; the revenues generated and market penetration gained from an advertising campaign; or the value of exercising an option at a given time. Heckman and Navarro (2007) unite and contribute to the literatures on dynamic discrete choice and dynamic treatment effects. For both classes of models, they present semiparametric identification analyses. We summarize their work in this section. It is formulated in discrete time, which facilitates the specification of richer unobserved and observed covariate processes than those entertained in the continuous-time framework of Abbring and Van den Berg (2003b).

Heckman and Navarro extend the literature on treatment effects to model choices of treatment times and the consequences of choice and link the literature on treatment effects to the literature on precisely formulated structural dynamic

discrete-choice models generated from index models crossing thresholds. They show the value of precisely formulated economic models in extracting the information sets of agents, in providing model identification, in generating the standard treatment effects and in enforcing the nonanticipating behavior condition (NA) discussed in Sect. 24.3.1.[46]

They establish the semiparametric identifiability of a class of dynamic discrete-choice models for stopping times and associated outcomes in which agents sequentially update the information on which they act. They also establish identifiability of a new class of reduced-form duration models that generalize conventional discrete-time duration models to produce frameworks with much richer time series properties for unobservables and general time-varying observables and patterns of duration dependence than conventional duration models. Their analysis of identification of these generalized models requires richer variation driven by observables than is needed in the analysis of the more restrictive conventional models. However, it does not require conventional period-by-period exclusion restrictions, which are often difficult to justify. Instead, they rely on curvature restrictions across the index functions generating the durations that can be motivated by dynamic economic theory.[47] Their methods can be applied to a variety of outcome measures including durations.

The key to their ability to identify structural models is that they supplement information on stopping times or time to treatment with additional information on measured consequences of choices of time to treatment as well as measurements. The dynamic discrete-choice literature surveyed in Rust (1994) and Magnac and Thesmar (2002) focuses on discrete-choice processes with general preferences and state vector evolution equations, typically Markovian in nature. Rust's 1994 paper contains negative results on nonparametric identification of discrete-choice processes. Magnac and Thesmar (2002) present some positive results on nonparametric identification if certain parameters or distributions of unobservables are assumed to be known. Heckman and Navarro (2007) produce positive results on nonparametric identification of a class of dynamic discrete-choice models based on expected income maximization developed in labor economics by Flinn and Heckman (1982), Keane and Wolpin (1997) and Eckstein and Wolpin (1999). These frameworks are dynamic versions of the Roy model. Heckman and Navarro (2007) show how use of cross-equation restrictions joined with data on supplementary measurement systems can undo Rust's nonidentification result. We exposit their work and the related literature in this section. With their structural framework, they can distinguish objective outcomes from subjective outcomes (valuations by the decision maker) in a dynamic setting. Applying their analysis to health economics, they can identify the causal effects on health of a medical treatment as well as the associated subjective

[46] Aakvik, Heckman, and Vytlacil (2005); Heckman, Tobias, and Vytlacil (2001, 2003); Carneiro, Hansen, and Heckman (2001, 2003) and Heckman and Vytlacil (2005) show how standard treatment effects can be generated from structural models.

[47] See Heckman and Honoré (1989) for examples of such an identification strategy in duration models. See also Cameron and Heckman (1998).

pain and suffering of a treatment regime for the patient.[48] Attrition decisions also convey information about agent preferences about treatment.[49]

They do not rely on the assumption of conditional independence of unobservables with outcomes, given observables, that is used throughout much of the dynamic discrete-choice literature and the dynamic treatment literature surveyed in Sect. 24.3.[50] As noted in Sect. 24.2, sequential conditional independence assumptions underlie recent work on reduced-form dynamic treatment effects.[51] The semiparametric analysis of Heckman and Navarro (2007) based on factors generalizes matching to a dynamic setting. In their paper, some of the variables that would produce conditional independence and would justify matching if they were observed, are treated as unobserved match variables. They are integrated out and their distributions are identified.[52] They consider two classes of models. We review both.

24.5.1 Semi-parametric Duration Models and Counterfactuals

Heckman and Navarro (2007), henceforth HN, develop a semiparametric index model for dynamic discrete choices that extends conventional discrete-time duration analysis. They separate out duration dependence from heterogeneity in a semiparametric framework more general than conventional discrete-time duration models. They produce a new class of reduced-form models for dynamic treatment effects by adjoining time-to-treatment outcomes to the duration model. This analysis builds on Heckman (1981a,b,c).

Their models are based on a latent variable for choice at time s,

$$I(s) = \Psi(s, Z(s)) - \eta(s) ,$$

where the $Z(s)$ are observables and $\eta(s)$ are unobservables from the point of view of the econometrician. Treatments at different times may have different outcome consequences which they model after analyzing the time to treatment equation. Define $D(s)$ as an indicator of receipt of treatment at date s. Treatment is taken the first time $I(s)$ becomes positive. Thus,

$$D(s) = \mathbf{1}[I(s) \geq 0, I(s-1) < 0, \ldots, I(1) < 0] ,$$

[48] See Chan and Hamilton (2006) for a structural dynamic empirical analysis of this problem.

[49] See Heckman and Smith (1998). Use of participation data to infer preferences about outcomes is developed in Heckman (1974).

[50] See, e.g., Rust (1987); Manski (1993); Hotz and Miller (1993) and the papers cited in Rust (1994).

[51] See, e.g., Gill and Robins (2001) and Lechner and Miquel (2002).

[52] For estimates based on this idea, see Carneiro et al. (2003); Aakvik et al. (2005); Cunha and Heckman (2007, 2008); Cunha, Heckman, and Navarro (2005, 2006); and Heckman and Navarro (2005).

where the indicator function $\mathbf{1}[\cdot]$ takes the value of 1 if the term inside the braces is true.[53] They derive conditions for identifying a model with general forms of duration dependence in the time to treatment equation using a large sample from the distribution of (D,Z).

24.5.1.1 Single-Spell Duration Model

Individuals are assumed to start spells in a given (exogenously determined) state and to exit the state at the beginning of time period S.[54] S is thus a random variable representing total completed spell length. Let $D(s) = 1$ if the individual exits at time s, $S = s$, and $D(s) = 0$ otherwise. In an analysis of drug treatments, S is the discrete-time period in the course of an illness at the beginning of which the drug is administered. Let $\bar{S}$ $(< \infty)$ be the upper limit on the time the agent being studied can be at risk for a treatment. It is possible in this example that $D(1) = 0, \ldots, D(\bar{S}) = 0$, so that a patient never receives treatment. In a schooling example, "treatment" is not schooling, but rather dropping out of schooling.[55] In this case, $\bar{S}$ is an upper limit to the number of years of schooling, and $D(\bar{S}) = 1$ if $D(1) = 0, \ldots, D(\bar{S}-1) = 0$.

The duration model can be specified recursively in terms of the threshold-crossing behavior of the sequence of underlying latent indices $I(s)$. Recall that $I(s) = \Psi(s, Z(s)) - \eta(s)$, with $Z(s)$ being the regressors that are observed by the analyst. The $Z(s)$ can include expectations of future outcomes given current information in the case of models with forward-looking behavior. For a given stopping time s, let $D^s = (D(1), \ldots, D(s))$ and designate by $d(s)$ and d^s values that $D(s)$ and D^s assume. Thus, $d(s)$ can be zero or one and d^s is a sequence of s zeros or a sequence containing $s-1$ zeros and a single one. Denote a sequence of all zeros by (0), regardless of its length. Then,

$$\begin{aligned} D(1) &= \mathbf{1}[I(1) \geq 0] \\ &\text{and} \\ D(s) &= \begin{cases} \mathbf{1}[I(s) \geq 0] & \text{if } D^{s-1} = (0) \\ 0 & \text{otherwise,} \end{cases} \qquad s = 2, \ldots, \bar{S}\,. \end{aligned} \tag{24.11}$$

For $s = 2, \ldots, \bar{S}$, the indicator $\mathbf{1}[I(s) \geq 0]$ is observed if and only if the agent is still at risk of treatment, $D^{s-1} = (0)$. To identify period s parameters from period s outcomes, one must condition on all past outcomes and control for any selection effects.

---

[53] This framework captures the essential feature of any stopping time model. For example, in a search model with one wage offer per period, $I(s)$ is the gap between market wages and reservation wages at time s. See, e.g., Flinn and Heckman (1982). This framework can also approximate the explicit dynamic discrete-choice model analyzed in Sect. 24.5.2.

[54] Thus we abstract from the initial-conditions problem discussed in Heckman (1981b).

[55] In the drug treatment example, S may designate the time a treatment regime is completed.

Let $Z = \big(Z(1),\ldots,Z(\bar{S})\big)$, and let $\eta = (\eta(1),\ldots,\eta(\bar{S}))$. Assume that Z is statistically independent of η. Heckman and Navarro (2007) assume that $\Psi(s,Z(s)) = Z(s)\gamma_s$. We deal with a more general case. $\Psi(Z) = \big(\Psi(1,Z(1)),\ldots,\Psi(\bar{S},Z(\bar{S}))\big)$. We let Ψ denote the abstract parameter. Depending on the values assumed by $\Psi(s,Z(s))$, one can generate very general forms of duration dependence that depend on the values assumed by the $Z(s)$. HN allow for period-specific effects of regressors on the latent indices generating choices.

This model is the reduced form of a general dynamic discrete-choice model. Like many reduced-form models, the link to choice theory is not clearly specified. It is not a conventional multinomial choice model in a static (perfect certainty) setting with associated outcomes.

24.5.1.2 Identification of Duration Models with General Error Structures and Duration Dependence

Heckman and Navarro (2007) establish semiparametric identification of the model of equation (24.11) assuming access to a large sample of i.i.d. (D,Z) observations. Let $Z^s = (Z(1),\ldots,Z(s))$. Data on (D,Z) directly identify the conditional probability $\Pr(D(s) = d(s)|Z^s,\ D^{s-1} = (0))$ *a.e.* $F_{Z^s|D^{s-1}=(0)}$ where $F_{Z^s|D^{s-1}=(0)}$ is the distribution of Z^s conditional on previous choices $D^{s-1} = (0)$. Assume that $(\Psi,F_\eta) \in \Phi \times \mathscr{H}$, where F_η is the distribution of η and $\Phi \times \mathscr{H}$ is the parameter space. The goal is to establish conditions under which knowledge of $\Pr(D(s) = d(s)|Z,D^{s-1} = (0))$ *a.e.* $F_{Z|D^{s-1}=(0)}$ allows the analyst to identify a unique element of $\Phi \times \mathscr{H}$. They use a limit strategy that allows them to recover the parameters by conditioning on large values of the indices of the preceding choices. This identification strategy is widely used in the analysis of discrete choice.[56]

They establish sufficient conditions for the identification of model (24.11). We prove the following more general result:

Theorem 24.1. *For the model defined by (24.11), assume the following conditions:*

(i) $\eta \perp\!\!\!\perp Z$.

(ii) η *is an absolutely continuous random variable on* $\mathbb{R}^{\bar{S}}$ *with support* $\prod_{s=1}^{\bar{S}}(\underline{\eta}(s), \overline{\eta}(s))$, *where* $-\infty \leq \underline{\eta}(s) < \overline{\eta}(s) \leq +\infty$, *for all* $s = 1,\ldots,\bar{S}$.

(iii) The $\Psi(s,Z(s))$ *satisfy the Matzkin (1992) conditions for identification of nonparametric binary choice models,* $s = 1,\ldots,\bar{S}$.[57]

(iv) $\text{Supp}\left(\Psi^{s-1}(Z),Z(s)\right) = \text{Supp}\left(\Psi^{s-1}(Z)\right) \times \text{Supp}\left(Z(s)\right)$, $s = 2,\ldots,\bar{S}$.

(v) $\text{Supp}(\Psi(Z)) \supseteq \text{Supp}(\eta)$.

[56] See, e.g., Manski (1988); Heckman (1990); Heckman and Honoré (1989, 1990); Matzkin (1992, 1993); Taber (2000); and Carneiro et al. (2003). A version of the strategy of this proof was first used in psychology where agent choice sets are eliminated by experimenter manipulation. The limit set argument effectively uses regressors to reduce the choice set confronting agents. See Falmagne (1985) for a discussion of models of choice in psychology.

[57] See Abbring and Heckman (2007, Appendix B.1) for a review of the conditions Matzkin (1992) imposes for identification of nonparametric binary choice models. See also Matzkin (1994).

Then F_η and $\Psi(Z)$ are identified, where the $\Psi(s,Z(s))$, $s=1,\ldots,\bar{S}$, are identified over the relevant support admitted by (ii).

Proof. We sketch the proof for $\bar{S}=2$. The result for general $\bar{S}$ follows by a recursive application of this argument. Consider the following three probabilities.

(a) $\Pr(D(1)=1 \mid Z=z) = \int_{\underline{\eta}(1)}^{\Psi(1,z(1))} f_{\eta(1)}(u)\,\mathrm{d}u$

(b) $\Pr(D(2)=1, D(1)=0 \mid Z=z) = \int_{\underline{\eta}(2)}^{\Psi(2,z(2))} \int_{\Psi(1,z(1))}^{\bar{\eta}(1)} f_{\eta(1),\eta(2)}(u_1,u_2)\,\mathrm{d}u_1\mathrm{d}u_2.$

(c) $\Pr(D(2)=0, D(1)=0 \mid Z=z) = \int_{\Psi(2,z(2))}^{\bar{\eta}(2)} \int_{\Psi(1,z(1))}^{\bar{\eta}(1)} f_{\eta(1),\eta(2)}(u_1,u_2)\,\mathrm{d}u_1\mathrm{d}u_2.$

The left-hand sides are observed from data on those who stop in period 1 (a); those who stop in period 2 (b); and those who terminate in the "0" state in period 2 (c). From Matzkin (1992), we can identify $\Psi(1,z(1))$ and $F_{\eta(1)}$ from (a). Using (b), we can fix $z(2)$ and vary $\Psi(1,z(1))$. From (iv) and (v), there exists a limit set $\widetilde{\mathscr{Z}_1}$, possibly dependent on $z(2)$, such that $\lim_{z(1)\to\widetilde{\mathscr{Z}_1}} \Psi(1,z(1)) = \underline{\eta}(1)$. Thus we can construct

$$\Pr(D(2)=0 \mid Z=z) = \int_{\Psi(2,z(2))}^{\bar{\eta}(2)} f_{\eta(2)}(u_2)\,\mathrm{d}u_2$$

and identify $\Psi(2,z(2))$ and $F_{\eta(2)}$. Using the $\Psi(1,z(1))$, $\Psi(2,z(2))$, one can trace out the joint distribution $F_{\eta(1),\eta(2)}$ over its support. Under the Matzkin conditions, identification is achieved on a non-negligible set. The proof generalizes in a straightforward way to general $\bar{S}$. ■

Observe that if the $\eta(s)$ are bounded by finite upper and lower limits, we can only determine the $\Psi(s,Z(s))$ over the limits so defined. Consider the first step of the proof. Under the Matzkin conditions, $F_{\eta(1)}$ is known. From assumption (ii), we can determine

$$\Psi(1,z(1)) = F_{\eta(1)}^{-1}(\Pr(D(1)=1 \mid Z=z))\,,$$

but only over the support $(\underline{\eta}(1),\bar{\eta}(1))$. If the support of $\eta(1)$ is $\mathbb{R}$, we determine $\Psi(1,z(1))$ for all $z(1)$. Heckman and Navarro (2007) analyze the special case $\Psi(s,Z(s)) = Z(s)\gamma_s$ and invoke sequential rank conditions to identify γ_s, even over limited supports. They also establish that the limit sets are non-negligible in this case so that standard definitions of identifiability (see, e.g., Matzkin, 1992) will be satisfied.[58] Construction of the limit set $\widetilde{Z}_s, s=1,\ldots,\bar{S}$, depends on the functional form specified for the $\Psi(s,Z(s))$. For the linear-in-parameters case $\Psi(s,Z(s)) = Z(s)\gamma_s$, they are obtained by letting arguments get big or small. Matzkin (1992) shows how to establish the limit sets for functions in her family of functions.

[58] Heckman and Navarro (2007) prove their theorem for a model where $D(s) = \mathbf{1}[I(s) \leq 0]$ if $D^{s-1} = (0), s=2,\ldots,\bar{S}$. Our formulation of their result is consistent with the notation in this chapter.

A version of Theorem 24.1 with $\Psi(s, Z(s)) = Z(s)\gamma_s$ that allows dependence between Z and η^s except for one component can be proved using the analysis of Lewbel (2000) and Honoré and Lewbel (2002).[59]

The assumptions of Theorem 24.1 will be satisfied if there are transition-specific exclusion restrictions for Z with the required properties. As noted in Sect. 24.4, in models with many periods, this may be a demanding requirement. Very often, the Z variables are time invariant and so cannot be used as exclusion restrictions. Corollary 1 in HN, for the special case $\Psi(s, Z(s)) = Z(s)\gamma_s$, tells us that the HN version of the model can be identified, even if there are no conventional exclusion restrictions and the $Z(s)$ are the *same* across all time periods, if sufficient structure is placed on how the γ_s vary with s. Variations in the values of γ_s across time periods arise naturally in finite horizon dynamic discrete-choice models where a shrinking horizon produces different effects of the same variable in different periods. For example, in Wolpin's (1987) analysis of a search model, the value function depends on time and the derived decision rules weight the same invariant characteristics differently in different periods. In a schooling model, parental background and resources may affect education continuation decisions differently at different stages of the schooling decision. The model generating (24.11) can be semiparametrically identified without transition-specific exclusions if the duration dependence is sufficiently general. For a proof, see Corollary 1 in Heckman and Navarro (2007).

The conditions of Theorem 24.1 are somewhat similar to the conditions on the regressor effects needed for identification of the continuous-time event-history models in Sect. 24.4. One difference is that the present analysis requires independent variation of the regressor effects over the support of the distribution of the unobservables generating outcomes. The continuous-time analysis based on the functional form of the mixed proportional hazard model (MPH) as analyzed by Abbring and Van den Berg (2003a) only requires local independent variation.

Theorem 24.1 and Corollary 1 in HN have important consequences. The $\Psi\ (s, Z(s))$, $s = 1, \ldots, \bar{S}$, can be interpreted as duration dependence parameters that are modified by the $Z(s)$ and that vary across the spell in a more general way than is permitted in mixed proportional hazards (MPH), generalized accelerated failure time (GAFT) models or standard discrete-time hazard models.[60] Duration dependence in conventional specifications of duration models is usually generated by variation in model intercepts. The regressors are allowed to interact with the duration dependence parameters. In the specifications justified by Theorem 24.1, the "heterogeneity" distribution F_η is identified for a general model. No special "permanent-transitory" structure is required for the unobservables although that specification is traditional in duration analysis. Their explicit treatment of the stochastic structure

[59] HN discuss a version of such an extension at their website. Lewbel's conditions are very strong. To account for general forms of dependence between Z and η^s requires modeling the exact form of the dependence. Nonparametric solutions to this problem remain an open question in the literature on dynamic discrete choice. One solution is to assume functional forms for the error terms, but in general, this is not enough to identify the model without further restrictions imposed. See Heckman and Honoré (1990).

[60] See Ridder (1990) for a discussion of these models.

of the duration model is what allows HN to link in a general way the unobservables generating the duration model to the unobservables generating the outcome equations that are introduced in the next section. Such an explicit link is not currently available in the literature on continuous-time duration models for treatment effects surveyed in Sect. 24.4, and is useful for modelling selection effects in outcomes across different treatment times. Their outcomes can be both discrete and continuous and are not restricted to be durations.

Under the conditions given in Corollary 1 of HN, no period-specific exclusion conditions are required on the Z. Hansen and Sargent (1980) and Abbring and Van den Berg (2003b) note that period-specific exclusions are not natural in reduced-form duration models designed to approximate forward-looking life cycle models. Agents make current decisions in light of their forecasts of future constraints and opportunities, and if they forecast some components well, and they affect current decisions, then they are in $Z(s)$ in period s. Corollary 1 in HN establishes identification without such exclusions. HN adjoin a system of counterfactual outcomes to their model of time to treatment to produce a model for dynamic counterfactuals. We summarize that work next.

24.5.1.3 Reduced-Form Dynamic Treatment Effects

This section reviews a reduced-form approach to generating dynamic counterfactuals developed by HN. They apply and extend the analysis of Carneiro et al. (2003) and Cunha et al. (2005, 2006) to generate *ex post* potential outcomes and their relationship with the time to treatment indices $I(s)$ analyzed in the preceding subsection. With reduced-form models, it is difficult to impose restrictions from economic theory or to make distinctions between *ex ante* and *ex post* outcomes. In the structural model developed below, these and other distinctions can be made easily.

Associated with each treatment time s, $s = 1,\ldots,\bar{S}$, is a vector of $\bar{T}$ outcomes,

$$Y(s,X,U(s)) = (Y(1,s,X,U(1,s)),\ldots,Y(t,s,X,U(t,s)),\ldots,Y(\bar{T},s,X,U(\bar{T},s))).$$

Outcomes depend on covariates X and $U(s) = (U(1,s),\ldots,U(t,s),\ldots,U(\bar{T},s))$ that are, respectively, observable and unobservable by the econometrician. Elements of $Y(s,X,U(s))$ are outcomes associated with stopping or receiving treatment at the beginning of period s. They are factual outcomes if treatment s is actually selected ($S = s$ and $D(s) = 1$). Outcomes corresponding to treatments s' that are not selected ($D(s') = 0$) are counterfactuals. The outcomes associated with each treatment may be different, and indeed the treatments administered at different times may be different.

The components $Y(t,s,X,U(t,s))$ of the vector $Y(s,X,U(s))$ can be interpreted as the outcomes revealed at age t, $t = 1,\ldots,\bar{T}$, and may themselves be vectors. The reduced-form approach presented in this section is not sufficiently rich to capture the notion that agents revise their anticipations of components of $Y(s,X,U(s))$,

$s = 1, \ldots, \bar{S}$, as they acquire information over time. This notion is systematically developed using the structural model discussed below in Sect. 24.5.2.

The treatment "times" may be stages that are not necessarily connected with real times. Thus s may be a schooling level. The correspondence between stages and times is exact if each stage takes one period to complete. Our notation is more flexible, and time and periods can be defined more generally. Our notation in this section accommodates both cases.

Henceforth, whenever we have random variables with multiple arguments $R_0(s, Q_0, \ldots)$ or $R_1(t, s, Q_0, \ldots)$ where the argument list begins with treatment state s or both age t and state s (perhaps followed by other arguments $Q_0, \ldots$), we will make use of several condensed notations: (a) dropping the first argument as we collect the components into vectors $R_0(Q_0, \ldots)$ or $R_1(s, Q_0, \ldots)$ of length $\bar{S}$ or $\bar{T}$, respectively, and (b) going further in the case of R_1, dropping the s argument as we collect the vectors $R_1(s, Q_0, \ldots)$ into a single $\bar{S} \times \bar{T}$ array $R_1(Q_0, \ldots)$, but also (c) suppressing one or more of the other arguments and writing $R_1(t, s)$ or $R_1(t, s, Q_0)$ instead of $R_1(t, s, Q_0, Q_1, \ldots)$, etc. This notation is sufficiently rich to represent the life cycle of outcomes for persons who receive treatment at s. Thus, in a schooling example, the components of this vector may include life cycle earnings, employment, and the like associated with a person with characteristics X, $U(s)$, $s = 1, \ldots, \bar{S}$, who completes s years of schooling and then forever ceases schooling. It could include earnings while in school at some level for persons who will eventually attain further schooling as well as post-school earnings.

We measure age and treatment time on the same time scale, with origin 1, and let $\bar{T} \geq \bar{S}$. Then, the $Y(t, s, X, U(t, s))$ for $t < s$ are outcomes realized while the person is in school at age t (s is the time the person will leave school; t is the current age) and before "treatment" (stopping schooling) has occurred. When $t \geq s$, these are post-school outcomes for treatment with s years of schooling. In this case, $t - s$ is years of post-school experience. In the case of a drug trial, the $Y(t, s, X, U(t, s))$ for $t < s$ are measurements observed before the drug is taken at s and if $t \geq s$, they are the post-treatment measurements.

Following Carneiro et al. (2003), the variables in $Y(t, s, X, U(t, s))$ may include discrete, continuous or mixed discrete-continuous components. For the discrete or mixed discrete-continuous cases, HN assume that latent continuous variables cross thresholds to generate the discrete components. Durations can be generated by latent index models associated with each outcome crossing thresholds analogous to the model presented in (24.11). In this framework, for example, we can model the effect of attaining s years of schooling on durations of unemployment or durations of employment.

The reduced-form analysis in this section does not impose restrictions on the temporal (age) structure of outcomes across treatment times in constructing outcomes and specifying identifying assumptions. Each treatment time can have its own age path of outcomes pre and post treatment. Outcomes prior to treatment and outcomes after treatment are treated symmetrically and both may be different for different treatment times. In particular, HN can allow earnings at age t for people who receive treatment at some future time s' to differ from earnings at age t for people

who receive treatment at some future time s'', $\min(s', s'') > t$ even after controlling for U and X.

This generality is in contrast with the analyses of Robins (1997) and Gill and Robins (2001) discussed in Sect. 24.3 and the analysis of Abbring and Van den Berg (2003b) discussed in Sect. 24.4. These analyses require exclusion of such anticipation effects to secure identification, because their models attribute dependence of treatment on past outcomes to selection effects. The sequential randomization assumption (M-1) underlying the work of Gill and Robins allows treatment decisions $S(t)$ at time t to depend on past outcomes $Y_{p_0}^{t-1}$ in a general way. Therefore, without additional restrictions, it is not possible to also identify *causal* (anticipatory) effects of treatment $S(t)$ on $Y_{p_0}^{t-1}$. The no-anticipation condition (NA) excludes such effects and secures identification in their framework.[61] It is essential for applying the conditional independence assumptions in deriving the g-computation formula.

HN's very different approach to identification allows them to incorporate anticipation effects. As in their analysis of the duration model, they assume that there is an exogenous source of independent variation of treatment decisions, independent of past outcomes. Any variation in current outcomes with variation in future treatment decisions induced by this exogenous source cannot be due to selection effects (since they explicitly control for the unobservables) and is interpreted as anticipatory effects of treatment in their framework. However, their structural analysis naturally excludes such effects (see Sect. 24.5.2 below). Therefore, a natural interpretation of the ability of HN to identify anticipatory effects is that they have overidentifying restrictions that allow them to test their model and, if necessary, relax their assumptions.

In a model with uncertainty, agents act on and value *ex ante* outcomes. The model developed below in Sect. 24.5.2 distinguishes *ex ante* from *ex post* outcomes. The model developed in this section cannot because, within it, it is difficult to specify the information sets on which agents act or the mechanism by which agents forecast and act on $Y(s, X, U(s))$ when they are making choices.

---

[61] The role of the no-anticipation assumption in Abbring and Van den Berg (2003b) is similar. However, their main analysis assumes an asymmetric treatment-outcome setup in which treatment is not observed if it takes place after the outcome transition. In that case, the treatment time is censored at the outcome time. In this asymmetric setup, anticipatory effects of treatment on outcomes cannot be identified because the econometrician cannot observe variation of outcome transitions with future treatment times. This point may appear to be unrelated to the present discussion, but it is not. As was pointed out by Abbring and Van den Berg (2003b), and in Sect. 24.4, the asymmetric Abbring and Van den Berg (2003b) model can be extended to a fully symmetric bivariate duration model in which treatment hazards may be causally affected by the past occurrence of an outcome event just like outcomes may be affected by past treatment events. This model could be used to analyze data in which both treatment and outcome times are fully observed. In this symmetric setup, any dependence in the data of the time-to-treatment hazard on past outcome events is interpreted as an effect of outcomes on future treatment decisions, and not an anticipatory effect of treatment on past outcomes. If one does not restrict the effects of outcomes on future treatment, without further restrictions, the data on treatments occurring after the outcome event carry no information on anticipatory effects of treatment on outcomes and they face an identification problem similar to that in the asymmetric case.

One justification for not making an *ex ante – ex post* distinction is that the agents being modeled operate under perfect foresight even though econometricians do not observe all of the information available to the agents. In this framework, the $U(s), s = 1, \ldots, \bar{S}$, are an ingredient of the econometric model that accounts for the asymmetry of information between the agent and the econometrician studying the agent.

Without imposing assumptions about the functional structure of the outcome equations, it is not possible to nonparametrically identify counterfactual outcome states $Y(s, X, U(s))$ that have never been observed. Thus, in a schooling example, HN assume that analysts observe life cycle outcomes for some persons for each stopping time (level of final grade completion) and our notation reflects this.[62] However, analysts do not observe $Y(s, X, U(s))$ for all s for anyone. A person can have only one stopping time (one completed schooling level). This observational limitation creates our evaluation problem, the "fundamental problem of causal inference".[63]

In addition to this problem, there is the standard selection problem that the $Y(s, X, U(s))$ are only observed for persons who stop at s and not for a random sample of the population. The selected distribution may not accurately characterize the population distribution of $Y(s, X, U(s))$ for persons selected at random. Note also that without further structure, we can only identify treatment responses within a given policy environment. In another policy environment, where the rules governing selection into treatment and/or the outcomes from treatment may be different, the same time to treatment may be associated with entirely different responses.[64] We now turn to the HN analysis of identification of outcome and treatment time distributions.

24.5.1.4 Identification of Outcome and Treatment Time Distributions

We assume access to a large i.i.d. sample from the distribution of $(S, Y(S, X, U(S)), X, Z)$, where S is the stopping time, X are the variables determining outcomes and Z are the variables determining choices. We also know $\Pr(S = s \mid Z = z)$, for $s = 1, \ldots, \bar{S}$, from the data. For expositional convenience, we first consider the case of scalar outcomes $Y(S, X, U(S))$. An analysis for vector $Y(S, X, U(S))$ is presented in HN and is discussed below.

Consider the analysis of continuous outcomes. HN analyze more general cases. Their results extend the analyses of Heckman and Honoré (1990); Heckman (1990) and Carneiro et al. (2003) by considering choices generated by a stopping time model. To simplify the notation in this section, assume that the scalar outcome associated with stopping at time s can be written as $Y(s) = \mu(s, X) + U(s)$, where $Y(s)$

[62] In practice, analysts can only observe a portion of the life cycle after treatment. See the discussion on pooling data across samples in Cunha et al. (2005) to replace missing life cycle data.

[63] See Holland (1986) or Gill and Robins (2001).

[64] This is the problem of general equilibrium effects, and leads to violation of the policy invariance conditions. See Heckman et al. (1998), Heckman et al. (1999) or Abbring and Van den Berg (2003b) for discussion of this problem.

is shorthand for $Y(s,X,U(s))$. $Y(s)$ is observed only if $D(s)=1$ where the $D(s)$ are generated by the model analyzed in Theorem 24.1. Write $I(s)=\Psi(s,Z(s))-\eta(s)$. Assume that the $\Psi(s,Z(s))$ belong to the Matzkin (1992) class of functions. We use the condensed representations I, $\Psi(Z)$, η, Y, $\mu(X)$ and U as described in the previous subsection.

Heckman and Navarro permit general stochastic dependence within the components of U, within the components of η and across the two vectors. They assume that (X,Z) are independent of (U,η). Each component of (U,η) has a zero mean. The joint distribution of (U,η) is assumed to be absolutely continuous.

With "sufficient variation" in the components of $\Psi(Z)$, one can identify $\mu(s,X)$, $[\Psi(1,Z(1)),\ldots,\Psi(s,Z(s))]$ and the joint distribution of $U(s)$ and η^s. This enables the analyst to identify average treatment effects across all stopping times, since one can extract $E(Y(s)-Y(s')\mid X=x)$ from the marginal distributions of $Y(s)$, $s=1,\ldots,\bar{S}$.

Theorem 24.2. *Write* $\Psi^s(Z)=(\Psi(1,Z(1),\ldots,\Psi(s,Z(s)))$. *Assume in addition to the conditions in Theorem 24.1 that*

(i) $E[U(s)]=0$. $(U(s),\eta^s)$ *are continuous random variables with support* $\mathrm{Supp}(U(s))\times\mathrm{Supp}(\eta^s)$ *with upper and lower limits* $(\overline{U}(s),\overline{\eta}^s)$ *and* $(\underline{U}(s),\underline{\eta}^s)$, *respectively,* $s=1,\ldots,\bar{S}$. *These conditions hold for each component of each subvector. The joint system is thus variation free for each component with respect to every other component.*

(ii) $(U(s),\eta^s)\perp\!\!\!\perp(X,Z)$, $s=1,\ldots,\bar{S}$ *(independence).*

(iii) $\mu(s,X)$ *is a continuous function,* $s=1,\ldots,\bar{S}$.

(iv) $\mathrm{Supp}(\Psi(Z),X)=\mathrm{Supp}(\Psi(Z))\times\mathrm{Supp}(X)$.

Then one can identify $\mu(s,X)$, $\Psi^s(Z)$, $F_{\eta^s,U(s)}$, $s=1,\ldots,\bar{S}$, *where* $\Psi(Z)$ *is identified over the support admitted by condition (ii) of Theorem 24.1.*

Proof. See Abbring and Heckman (2007), Appendix C.

The proof in Abbring and Heckman (2007, Appendix C) covers the case of vector $Y(s,X,U(s))$ where each component is a continuous random variable. Appendix D of Abbring and Heckman (2007) states and proves a more general theorem for age-specific outcomes $Y(t,s,X,U(t,s))$, $t=1,\ldots,\bar{T}$, where Y can be a vector of continuous and discrete outcomes. In particular, HN can identify age-specific earnings flows associated with multiple sources of income.

Theorem 24.2 does not identify the joint distribution of $Y(1),\ldots,Y(\bar{S})$ because analysts observe only one of these outcomes for any person. Observe that exclusion restrictions in the arguments of the choice of treatment equation are not required to identify the counterfactuals. What is required is independent variation of arguments which might be achieved by exclusion conditions but can be obtained by other functional restrictions (see HN, Corollary 1, for example). One can identify the $\mu(s,X)$ (up to constants) without the limit set argument. Thus, one can identify certain features of the model without using the limit set argument. See HN.

As a by-product of Theorem 24.2, one can construct various counterfactual distributions of $Y(s)$ for agents with index crossing histories such that $D(s)=0$ (that is,

for whom $Y(s)$ is not observed). Define $B(s) = \mathbf{1}[I(s) \geq 0]$, $B^s = (B(1), \ldots, B(s))$, and let b^s denote a vector of possible values of B^s. $D(s)$ was defined as $B(s)$ if $B^{s-1} = (0)$ and 0 otherwise. Theorem 24.2 gives conditions under which the counterfactual distribution of $Y(s)$ for those with $D(s') = 1$, $s' \neq s$, can be constructed. More generally, it can be used to construct

$$\Pr\left(Y(s) \leq y(s) \mid B^{s'} = b^{s'}, X = x, Z = z\right)$$

for all of the $2^{s'}$ possible sequences $b^{s'}$ of $B^{s'}$ outcomes up to $s' \leq s$. If $b^{s'}$ equals a sequence of $s'-1$ zeros followed by a one, then $B^{s'} = b^{s'}$ corresponds to $D(s') = 1$. The event $B^{s'} = (0)$ corresponds to $D^{s'} = (0)$, i.e., $S > s'$. For all other sequences $b^{s'}$, $B^{s'} = b^{s'}$ defines a subpopulation of the agents with $D(s'') = 1$ for some $s'' < s'$ and multiple index crossings. For example, $B^{s'} = (0,1,0)$ corresponds to $D(2) = 1$ and $I(3) < 0$. This defines a subpopulation that takes treatment at time 2, but that would not take treatment at time 3 if it would not have taken treatment at time 2.[65] It is tempting to interpret such sequences with multiple crossings as corresponding to multiple entry into and exit from treatment. However, this is inconsistent with the stopping time model (24.11), and would require extension of the model to deal with recurrent treatment. Whether a threshold-crossing model corresponds to a structural model of treatment choice is yet another issue, which is taken up in the next section and is also addressed in Cunha, Heckman, and Navarro (2007).

The counterfactuals that are identified by fixing $D(s') = 1$ for different treatment times s' in the general model of HN have an asymmetric aspect. HN can generate $Y(s)$ distributions for persons who are treated at s or before. Without further structure, they cannot generate the distributions of these random variables for people who receive treatment at times after s.

The source of this asymmetry is the generality of duration model (24.11). At each stopping time s, HN acquire a new random variable $\eta(s)$ which can have arbitrary dependence with $Y(s)$ and $Y(s')$ for all s and s'. From Theorem 24.2, HN can identify the dependence between $\eta(s')$ and $Y(s)$ if $s' \leq s$. They cannot identify the dependence between $\eta(s')$ and $Y(s)$ for $s' > s$ without imposing further structure on the unobservables.[66] Thus, one can identify the distribution of college outcomes for high school graduates who do not go on to college and can compare these to outcomes for high school graduates, so they can identify the parameter "treatment on the untreated." However, one cannot identify the distribution of high school outcomes for college graduates (and hence treatment on the treated parameters) without imposing further structure.[67] Since one can identify the marginal distributions under

[65] Cunha et al. (2007) develop an ordered choice model with stochastic thresholds.

[66] One possible structure is a factor model which is applied to this problem in the next section.

[67] In the schooling example, one can identify treatment on the treated for the final category $\bar{S}$ since $D^{\bar{S}-1} = (0)$ implies $D(\bar{S}) = 1$. Thus at stage $\bar{S}-1$, one can identify the distribution of $Y(\bar{S}-1)$ for persons for whom $D(0) = 0, \ldots, D(\bar{S}-1) = 0, D(\bar{S}) = 1$. Hence, if college is the terminal state, and high school the state preceding college, one can identify the distribution of high school outcomes for college graduates.

the conditions of Theorem 24.2, one can identify pairwise average treatment effects for all s, s'.

It is interesting to contrast the model identified by Theorem 24.2 with a conventional static multinomial discrete-choice model with an associated system of counterfactuals, as presented in Heckman and Vytlacil (2007a, Appendix B) and analyzed in Abbring and Heckman (2007, Sect. 2). Using standard tools, it is possible to establish semiparametric identification of the conventional static model of discrete choice joined with counterfactuals and to identify all of the standard mean counterfactuals. For that model there is a fixed set of unobservables governing all choices of states. Thus the analyst does not acquire new unobservables associated with each stopping time as occurs in a dynamic model. In a dynamic model, selection effects for $Y(s)$ depend on the unobservables up to s but not later innovations Selection effects in a static discrete-choice model depend on a fixed set of unobservables for all outcomes. With suitable normalizations, HN identify the joint distributions of choices and associated outcomes without the difficulties, just noted, that appear in the reduced-form dynamic model. HN develop models for discrete outcomes including duration models.

24.5.1.5 Using Factor Models to Identify Joint Distributions of Counterfactuals

From Theorem 24.2 and its generalizations reported in HN, one can identify joint distributions of outcomes for each treatment time s and the index generating treatment times. One cannot identify the joint distributions of outcomes across treatment times. Moreover, as just discussed, one cannot, in general, identify treatment on the treated parameters.

Aakvik et al. (2005) and Carneiro et al. (2003) show how to use factor models to identify the joint distributions across treatment times and recover the standard treatment parameters. HN use their approach to identify the joint distribution of $Y = (Y(1), \ldots, Y(\bar{S}))$.

The basic idea underlying this approach is to use joint distributions for outcomes measured at each treatment time s along with the choice index to construct the joint distribution of outcomes across treatment choices. To illustrate how to implement this intuition, suppose that we augment Theorem 24.2 by appealing to Theorem 2 in Carneiro et al. (2003) to identify the joint distribution of the vector of outcomes at each stopping time along with $I^s = (I(1), \ldots, I(s))$ for each s. For each s, we may write

$$\begin{aligned} Y(t,s,X,U(t,s)) &= \mu(t,s,X) + U(t,s), \quad t = 1, \ldots, \bar{T} \\ I(s) &= \Psi(s, Z(s)) - \eta(s). \end{aligned}$$

The scale of $\Psi(s, Z(s))$ is determined from the Matzkin (1992) conditions. If we specify the Matzkin functions only up to scale, we determine the functions up to

scale and make a normalization. From Theorem 24.2, we can identify the joint distribution of $(\eta(1),\ldots,\eta(s),U(1,s),\ldots,U(\bar{T},s))$.

To review these concepts and their application to the model discussed in this section, suppose that we adopt a one-factor model where θ is the factor. It has mean zero. The errors can be represented by

$$\begin{aligned}\eta(s) &= \varphi_s\theta+\varepsilon_{\eta(s)}\\ U(t,s) &= \alpha_{t,s}\theta+\varepsilon_{t,s}, \qquad t=1,\ldots,\bar{T}, \qquad s=1,\ldots,\bar{S}\,.\end{aligned}$$

The θ are independent of all of the $\varepsilon_{\eta(s)}$, $\varepsilon_{t,s}$ and the ε's are mutually independent mean zero disturbances. The φ_s and $\alpha_{t,s}$ are factor loadings. Since θ is an unobservable, its scale is unknown. One can set the scale of θ by normalizing one factor loading, say $\alpha_{\bar{T},\bar{S}}=1$. From the joint distribution of $(\eta^s,U(s))$, one can identify σ_θ^2, $\alpha_{t,s},\varphi_s,t=1,\ldots,\bar{T}$, for $s=1,\ldots,\bar{S}$, using the arguments presented in, e.g., Abbring and Heckman (2007, Sect. 2.8). A sufficient condition is $\bar{T}\geq 3$, but this ignores possible additional information from cross-system restrictions. From this information, one can form for $t\neq t'$ or $s\neq s''$ or both,

$$\operatorname{Cov}\left(U(t,s),U\left(t',s''\right)\right)=\alpha_{t,s}\alpha_{t',s''}\sigma_\theta^2\,,$$

even though the analyst does not observe outcomes for the same person at two different stopping times. In fact, one can construct the joint distribution of $(U,\eta)=(U(1),\ldots,U\left(\bar{S}\right),\eta)$. From this joint distribution, one can recover the standard mean treatment effects as well as the joint distributions of the potential outcomes. One can determine the percentage of participants at treatment time s who benefit from participation compared to what their outcomes would be at other treatment times. One can perform a parallel analysis for models for discrete outcomes and durations. The analysis can be generalized to multiple factors. Conventional factor analysis assumes that the unobservables are normally distributed. Carneiro et al. (2003) establish nonparametric identifiability of the θ's and the ε's and their analysis of nonparametric identifiability applies here.

Theorem 24.2, strictly applied, actually produces only one scalar outcome along with one or more choices for each stopping time.[68] If vector outcomes are not available, access to a measurement system M that assumes the same values for each stopping time can substitute for the need for vector outcomes for Y. Let M_j be the jth component of this measurement system. Write

$$M_j=\mu_{j,M}(X)+U_{j,M},\quad j=1,\ldots,J\,,$$

where $U_{j,M}$ are mean zero and independent of X.

Suppose that the $U_{j,M}$ have a one-factor structure so $U_{j,M}=\alpha_{j,M}\theta+\varepsilon_{j,M}$, $j=1,\ldots,J$, where the $\varepsilon_{j,M}$ are mean zero, mutually independent random variables, independent of the θ. Adjoining these measurements to the one outcome measure $Y(s)$ can substitute for the measurements of $Y(t,s)$ used in the previous example.

[68] HN and Abbring and Heckman (2007) analyze the vector-outcome case.

In an analysis of schooling, the M_j can be test scores that depend on ability θ. Ability is assumed to affect outcomes $Y(s)$ and the choice of treatment times indices.

The factor models implement a matching on unobservables assumption, $\{Y(s)\}_{s=1}^{\bar{S}} \perp\!\!\!\perp S \mid X, Z, \theta$. HN allow for the θ to be unobserved variables and present conditions under which their distributions can be identified.

24.5.1.6 Summary of the Reduced-Form Model

A limitation of the reduced-form approach pursued in this section is that, because the underlying model of choice is not clearly specified, it is not possible without further structure to form, or even define, the marginal treatment effect analyzed in Heckman and Vytlacil (1999, 2001, 2005, 2007a,b) or Heckman, Urzua, and Vytlacil (2006). The absence of well defined choice equations is problematic for the models analyzed thus far in this section of our chapter, although it is typical of many statistical treatment effect analyses.[69] In this framework, it is not possible to distinguish objective outcomes from subjective evaluations of outcomes, and to distinguish *ex ante* from *ex post* outcomes. Another limitation of this analysis is its strong reliance on large support conditions on the regressors coupled with independence assumptions. Independence can be relaxed following Lewbel (2000) and Honoré and Lewbel (2002). The large support assumption plays a fundamental role here and throughout the entire evaluation literature.

HN develop an explicit economic model for dynamic treatment effects that allows analysts to make these and other distinctions. They extend the analysis presented in this subsection to a more precisely formulated economic model. They explicitly allow for agent updating of information sets. A well posed economic model enables economists to evaluate policies in one environment and accurately project them to new environments as well as to accurately forecast new policies never previously experienced. We now turn to an analysis of a more fully articulated structural econometric model.

24.5.2 A Sequential Structural Model with Option Values

This section analyzes the identifiability of a structural sequential optimal stopping time model. HN use ingredients assembled in the previous sections to build an economically interpretable framework for analyzing dynamic treatment effects. For specificity, HN focus on a schooling model with associated earnings outcomes that is motivated by the research of Keane and Wolpin (1997) and Eckstein and Wolpin

[69] Heckman (2005) and Heckman and Vytlacil (2007a,b) point out that one distinctive feature of the economic approach to program evaluation is the use of choice theory to define parameters and evaluate alternative estimators.

(1999). They explicitly model costs and build a dynamic version of a Roy model. We briefly survey the literature on dynamic discrete choice in Sect. 24.5.5 below.

In the model of this section, it is possible to interpret the literature on dynamic treatment effects within the context of an economic model; to allow for earnings while in treatment as well as grade-specific tuition costs; to separately identify returns and costs; to distinguish private evaluations from "objective" *ex ante* and *ex post* outcomes and to identify persons at various margins of choice. In the context of medical economics, HN consider how to identify the pain and suffering associated with a treatment as well as the distribution of benefits from the intervention. They also model how anticipations about potential future outcomes associated with various choices evolve over the life cycle as sequential treatment choices are made.

In contrast to the analysis of Sect. 24.5.1, the identification proof for their dynamic choice model works in reverse starting from the last period and sequentially proceeding backward. This approach is required by the forward-looking nature of dynamic choice analysis and makes an interesting contrast with the analysis of identification for the reduced-form models which proceeds forward from initial period values.

HN use limit set arguments to identify the parameters of outcome and measurement systems for each stopping time $s = 1, \ldots, \bar{S}$, including means and joint distributions of unobservables. These systems are identified without invoking any special assumptions about the structure of model unobservables. When they invoke factor structure assumptions for the unobservables, they identify the factor loadings associated with the measurements (as defined in Sect. 24.5.1.5) and outcomes. They also nonparametrically identify the distributions of the factors and the distributions of the innovations to the factors. With the joint distributions of outcomes and measurements in hand for each treatment time, HN can identify cost (and preference) information from choice equations that depend on outcomes and costs (preferences). HN can also identify joint distributions of outcomes across stopping times. Thus, they can identify the proportion of people who benefit from treatment. Their analysis generalizes the one shot decision models of Cunha and Heckman (2007, 2008); Cunha et al. (2005, 2006) to a sequential setting.

All agents start with 1 year of schooling at age 1 and then sequentially choose, at each subsequent age, whether to continue for another year in school. New information arrives at each age. One of the benefits of staying in school is the arrival of new information about returns. Each year of schooling takes 1 year of age to complete. There is no grade repetition. Once persons leave school, they never return.[70] As a consequence, an agent's schooling level equals her age up to the time $S \leq \bar{S}$ she leaves school. After that, ageing continues up to age $\bar{T} \geq \bar{S}$, but schooling does not. We again denote $D(s) = \mathbf{1}(S = s)$ for all $s \in \{1, \ldots, \bar{S}\}$. Let $\delta(t) = 1$ if a person has left school at or before age t; $\delta(t) = 0$ if a person is still in school.

A person's earnings at age t depend on her *current* schooling level s and whether she has left school on or before age t ($\delta(t) = 1$) or not ($\delta(t) = 0$). Thus,

[70] It would be better to derive such stopping behavior as a feature of a more general model with possible recurrence of states. Cunha et al. (2007) develop general conditions under which it is optimal to stop and never return.

$$Y(t,s,\delta(t),X) = \mu(t,s,\delta(t),X) + U(t,s,\delta(t)) \ . \tag{24.12}$$

Note that $Y(t,s,0,X)$ is only meaningfully defined if $s = t$, in which case it denotes the earnings of a person as a student at age and schooling level s. More precisely, $Y(s,s,0,X)$ denotes the earnings of an individual with characteristics X who is still enrolled in school at age and schooling level s and goes on to complete at least $s+1$ years of schooling. The fact that earnings in school depend only on the current schooling level, and not on the final schooling level obtained, reflects the no-anticipation condition (NA). $U(t,s,\delta(t))$ is a mean zero shock that is unobserved by the econometrician but may, or may not, be observed by the agent. $Y(t,s,1,X)$ is meaningfully defined only if $s \leq t$, in which case it denotes the earnings at age t of an agent who has decided to stop schooling at s.

The direct cost of remaining enrolled in school at age and schooling level s is

$$C(s,X,Z(s)) = \Phi(s,X,Z(s)) + W(s)$$

where X and $Z(s)$ are vectors of observed characteristics (from the point of view of the econometrician) that affect costs at schooling level s, and $W(s)$ are mean zero shocks that are unobserved by the econometrician that may or may not be observed by the agent. Costs are paid in the period before schooling is undertaken. The agent is assumed to know the costs of making schooling decisions at each transition. The agent is also assumed to know the X and $Z = (Z(1),\ldots,Z(\bar{S}-1))$ from age 1.[71]

The optimal schooling decision involves comparisons of the value of continuing in school for another year and the value of leaving school forever at each age and schooling level $s \in \{1,\ldots,\bar{S}-1\}$. We can solve for these values, and the optimal schooling decision, by backward recursion.

The agent's expected reward of stopping schooling forever at level and age s (i.e., receiving treatment s) is given by the expected present value of her remaining lifetime earnings:

$$R(s,I_s) = E\left(\sum_{j=0}^{\bar{T}-s} \left(\frac{1}{1+r} \right)^j Y(s+j,s,1,X) \,\middle|\, I_s \right) , \tag{24.13}$$

where I_s are the state variables generating the age-s-specific information set $\mathscr{I}_s$.[72] They include the schooling level attained at age s, the covariates X and Z, as well as all other variables known to the agent and used in forecasting future variables. Assume a fixed, nonstochastic, interest rate r.[73] The continuation value at age and schooling level s given information I_s is denoted by $K(s,I_s)$.

[71] These assumptions can be relaxed and are made for convenience. See Carneiro, Hansen, and Heckman (2003), Cunha, Heckman, and Navarro (2005) and Cunha and Heckman (2007) for a discussion of selecting variables in the agent's information set.

[72] We only consider the agent's information set here, and drop the subscript A for notational convenience.

[73] This assumption is relaxed in HN who present conditions under which r can be identified.

At $\bar{S}-1$, when an individual decides whether to stop or continue on to $\bar{S}$, the expected reward from remaining enrolled and continuing to $\bar{S}$ (i.e., the continuation value) is the earnings while in school less costs plus the expected discounted future return that arises from completing $\bar{S}$ years of schooling:

$$K\left(\bar{S}-1,I_{\bar{S}-1}\right)=Y\left(\bar{S}-1,\bar{S}-1,0,X\right)-C\left(\bar{S}-1,X,Z\left(\bar{S}-1\right)\right)$$
$$+\frac{1}{1+r}E\left(R\left(\bar{S},I_{\bar{S}}\right)\mid I_{\bar{S}-1}\right)$$

where $C\left(\bar{S}-1,X,Z\left(\bar{S}-1\right)\right)$ is the direct cost of schooling for the transition to $\bar{S}$. This expression embodies the assumption that each year of school takes 1 year of age. $I_{\bar{S}-1}$ incorporates all of the information known to the agent.

The value of being in school just before deciding on continuation at age and schooling level $\bar{S}-1$ is the larger of the two expected rewards that arise from stopping at $\bar{S}-1$ or continuing one more period to $\bar{S}$:

$$V\left(\bar{S}-1,I_{\bar{S}-1}\right)=\max\left\{R\left(\bar{S}-1,I_{\bar{S}-1}\right),K\left(\bar{S}-1,I_{\bar{S}-1}\right)\right\}.$$

More generally, at age and schooling level s, this value is

$$V\left(s,I_s\right)=\max\left\{R\left(s,I_s\right),K\left(s,I_s\right)\right\}$$
$$=\max\left\{R\left(s,I_s\right),\left(\begin{array}{c}Y\left(s,s,0,X\right)-C\left(s,X,Z\left(s\right)\right)\\+\frac{1}{1+r}E\left(V\left(s+1,I_{s+1}\right)\mid I_s\right)\end{array}\right)\right\}.^{74}$$

Following the exposition of the reduced-form decision rule in Sect. 24.5.1, define the decision rule in terms of a first passage of the "index" $R(s,I_s)-K(s,I_s)$,

$$D\left(s\right)=\mathbf{1}\left[R(s,I_s)-K(s,I_s)\geq 0,\right.$$
$$\left.R(s-1,I_{s-1})-K(s-1,I_{s-1})<0,\ldots,R(1,I_1)-K(1,I_1)<0\right].$$

An individual stops at the schooling level at the first age where this index becomes positive. From data on stopping times, one can nonparametrically identify

[74] This model allows no recall and is clearly a simplification of a more general model of schooling with option values. Instead of imposing the requirement that once a student drops out the student never returns, it would be useful to derive this property as a feature of the economic environment and the characteristics of individuals. Cunha et al. (2007) develop such conditions. In a more general model, different persons could drop out and return to school at different times as information sets are revised. This would create further option value beyond the option value developed in the text that arises from the possibility that persons who attain a given schooling level can attend the next schooling level in any future period. Implicit in this analysis of option values is the additional assumption that persons must work at the highest level of education for which they are trained. An alternative model allows individuals to work each period at the highest wage across all levels of schooling that they have attained. Such a model may be too extreme because it ignores the costs of switching jobs, especially at the higher educational levels where there may be a lot of job-specific human capital for each schooling level. A model with these additional features is presented in Heckman et al. (2007).

the conditional probability of stopping at s,

$$\Pr(S=s \mid X,Z) = \Pr\left(\begin{array}{l} R(s,I_s) - K(s,I_s) \geq 0, \\ R(s-1,I_{s-1}) - K(s-1,I_{s-1}) < 0, \ldots, \\ R(1,I_1) - K(1,I_1) < 0 \end{array} \middle| \; X,Z \right) .$$

HN use factor structure models based on the θ introduced in Sect. 24.5.1 to define the information updating structure. Agents learn about different components of θ as they evolve through life. The HN assumptions allow for the possibility that agents may know some or all the elements of θ at a given age t regardless of whether or not they determine earnings at or before age t. Once known, they are not forgotten. As agents accumulate information, they revise their forecasts of their future earnings prospects at subsequent stages of the decision process. This affects their decision rules and subsequent choices. Thus HN allow for learning which can affect both pretreatment outcomes and posttreatment outcomes.[75,76] All dynamic discrete-choice models make some assumptions about the updating of information and any rigorous identification analysis of this class of models must test among competing specifications of information updating.

Variables unknown to the agent are integrated out by the agent in forming expectations over future outcomes. Variables known to the agent are treated as constants by the agents. They are integrated out by the econometrician to control for heterogeneity. These are separate operations except for special cases. In general, the econometrician knows less than what the agent knows. The econometrician seeks to identify the distributions of the variables in the agent information sets that are used by the agents to form their expectations as well as the distributions of variables known to the agent and treated as certain quantities by the agent but not known by the econometrician. Determining which elements belong in the agent's information set can be done using the methods exposited in Cunha et al. (2005) and Cunha and Heckman (2007) who consider testing what components of X,Z,ε as well as θ are in the agent's information set. We briefly discuss this issue at the end of the next section.[77] HN establish semiparametric identification of the model assuming a given information structure. Determining the appropriate information structure

[75] This type of learning about unobservables can be captured by HN's reduced-form model, but not by Abbring and Van den Berg's (2003b) single-spell mixed proportional hazards model. Their model does not allow for time-varying unobservables. Abbring and Van den Berg develop a multiple-spell model that allows for time-varying unobservables. Moreover, their nonparametric discussion of (NA) and randomization does not exclude the sequential revelation to the agent of a finite number of unobserved factors although they do not systematically develop such a model.

[76] It is fruitful to distinguish models with exogenous arrival of information (so that information arrives at each age t independent of any actions taken by the agent) from information that arrives as a result of choices by the agent. The HN model is in the first class. The models of Miller (1984) or Pakes (1986) are in the second class. See our discussion in Sect. 24.5.5.

[77] The HN model of learning is clearly very barebones. Information arrives exogenously across ages. In the factor model, all agents who advance to a stage get information about additional factors at that stage of their life cycles but the realizations of the factors may differ across persons.

facing the agent and its evolution is an essential aspect of identifying any dynamic discrete-choice model.

Observe that agents with the same information variables I_t at age t have the same expectations of future returns, and the same continuation and stopping values. They make the same investment choices. Persons with the same *ex ante* reward, state and preference variables have the same *ex ante* distributions of stopping times. *Ex post*, stopping times may differ among agents with identical *ex ante* information. Controlling for I_t, future realizations of stopping times do not affect past rewards. This rules out the problem that the future can cause the past, which may happen in HN's reduced-form model. It enforces the (NA) condition of Abbring and Van den Berg. Failure to accurately model I_t produces failure of (NA).

HN establish semiparametric identification of their model without period-by-period exclusion restrictions. Their analysis extends Theorems 24.1 and 24.2 to an explicit choice-theoretic setting. They use limit set arguments to identify the joint distributions of earnings (for each treatment time s across t) and any associated measurements that do not depend on the stopping time chosen. For each stopping time, they construct the means of earnings outcomes at each age and of the measurements and the joint distributions of the unobservables for earnings and measurements. Factor analyzing the joint distributions of the unobservables, under conditions specified in Carneiro et al. (2003), they identify the factor loadings, and nonparametrically identify the distributions of the factors and the independent components of the error terms in the earnings and measurement equations. Armed with this knowledge, they use choice data to identify the distribution of the components of the cost functions that are not directly observed. They construct the joint distributions of outcomes across stopping times. They also present conditions under which the interest rate r is identified.

In their model, analysts can distinguish period by period *ex ante* expected returns from *ex post* realizations by applying the analysis of Cunha et al. (2005). See the survey in Heckman, Lochner, and Todd (2006) and Sect. 2 of Abbring and Heckman (2007) for discussions of this approach. Because they link choices to outcomes through the factor structure assumption, they can also distinguish *ex ante* preference or cost parameters from their *ex post* realizations. *Ex ante*, agents may not know some components of θ. *Ex post*, they do. All of the information about future rewards and returns is embodied in the information set $\mathscr{I}_t$. Unless the time of treatment is known with perfect certainty, it cannot cause outcomes prior to its realization.

The analysis of HN is predicated on specification of agent information sets. These information sets should be carefully distinguished from those of the econometrician. Cunha et al. (2005) present methods for determining which components of future outcomes are in the information sets of agents at each age, $\mathscr{I}_t$. If there are components unknown to the agent at age t, under rational expectations, agents form their value functions used to make schooling choices by integrating out the unknown components using the distributions of the variables in their information sets. Components that are known to the agent are treated as constants by the individual in forming the value function but as unknown variables by the econometrician and

their distribution is estimated. The true information set of the agent is determined from the set of possible specifications of the information sets of agents by picking the specification that best fits the data on choices and outcomes penalizing for parameter estimation. If neither the agent nor the econometrician knows a variable, the econometrician identifies the determinants of the distribution of the unknown variables that is used by the agent to form expectations. If the agent knows some variables, but the econometrician does not, the econometrician seeks to identify the distribution of the variables, but the agent treats the variables as known constants.

HN can identify all of the treatment parameters including the pairwise average treatment effect (ATE), the marginal treatment effect (MTE) for each transition (obtained by finding mean outcomes for individuals indifferent between transitions), all of the treatment on the treated and treatment on the untreated parameters and the population distribution of treatment effects by applying the analysis of Carneiro et al. (2003) and Cunha et al. (2005) to this model. Their analysis can be generalized to cover the case where there are vectors of contemporaneous outcome measures for different stopping times. See HN for proofs and details.[78]

24.5.3 Identification at Infinity

Heckman and Navarro (2007), and many other researchers, rely on identification at infinity to obtain their main identification results. Identification at infinity is required to identify the average treatment effect (ATE) using IV and control function methods and in the reduced-form discrete-time models developed in the previous subsections. While this approach is controversial, it is also testable. In any sample, one can plot the distributions of the probability of each state (exit time) to determine if the identification conditions are satisfied in any sample. Figure 24.1, presented by HN from the research of Heckman, Stixrud, and Urzua (2006), shows such plots for a six-state static schooling model that they estimate. To identify the marginal outcome distributions for each state, the support of the state probabilities should be the full unit interval. The identification at infinity condition is clearly not satisfied in their data.[79] Only the empirical distribution of the state probability of graduating from a 4-year college comes even close to covering the full unit interval. Thus, their empirical results rely on parametric assumptions, and ATE and the marginal distributions of outcomes are nonparametrically nonidentified in their data without invoking additional structure.

[78] The same limitations regarding independence assumptions between the regressors and errors discussed in the analysis of reduced forms apply to the structural model.

[79] One can always argue that they are satisfied in an infinite sample that has not yet been realized. That statement has no empirical content.

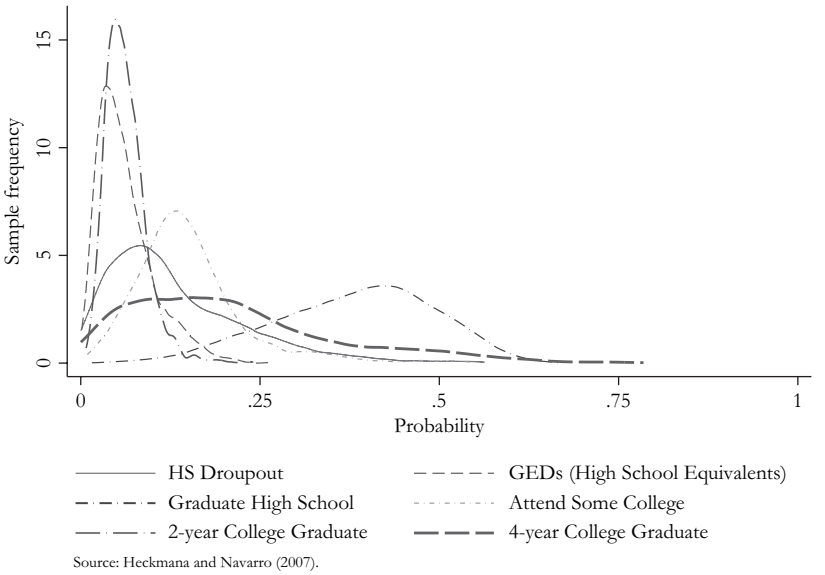

Fig. 24.1 Sample distribution of schooling attainment probabilities for males from the National Longitudinal Survey of Youth

24.5.4 Comparing Reduced-Form and Structural Models

The reduced-form model analyzed in Sect. 24.5.1 is typical of many reduced-form statistical approaches within which it is difficult to make important conceptual distinctions. Because agent choice equations are not modeled explicitly, it is hard to use such frameworks to formally analyze the decision makers' expectations, costs of treatment, the arrival of information, the content of agent information sets and the consequences of the arrival of information for decisions regarding time to treatment as well as outcomes. Key behavioral assumptions are buried in statistical assumptions. It is difficult to distinguish *ex post* from *ex ante* valuations of outcomes in the reduced-form models. Cunha et al. (2005); Carneiro et al. (2003) and Cunha and Heckman (2007, 2008) present analyses that distinguish *ex ante* anticipations from *ex post* realizations.[80] In reduced-form models, it is difficult to make the distinction between private evaluations and preferences (e.g., "costs" as defined in this section) from objective outcomes (the Y variables).

Statistical and reduced-form econometric approaches to analyzing dynamic counterfactuals appeal to uncertainty to motivate the stochastic structure of models. They do not explicitly characterize how agents respond to uncertainty or make treatment choices based on the arrival of new information (see Robins, 1989, 1997; Lok, 2007; Gill and Robins, 2001; Abbring and Van den Berg, 2003b; and Van der Laan and Robins, 2003). The structural approach surveyed in Sect. 24.5.2 and developed by

[80] See the summary of this literature in Heckman et al. (2006).

HN allows for a clear treatment of the arrival of information, agent expectations, and the effects of new information on choice and its consequences. In an environment of imperfect certainty about the future, it rules out the possibility of the future causing the past once the effects of agent information are controlled for.

The structural model developed by HN allows agents to learn about new factors (components of θ) as they proceed sequentially through their life cycles. It also allows agents to learn about other components of the model (see Cunha et al., 2005). Agent anticipations of when they will stop and the consequences of alternative stopping times can be sequentially revised. Agent anticipated payoffs and stopping times are sequentially revised as new information becomes available. The mechanism by which agents revise their anticipations is modeled and identified. See Cunha et al. (2005, 2006); Cunha and Heckman (2007, 2008) and Abbring and Heckman (2007) for further discussion of these issues and Heckman et al. (2006) for a partial survey of recent developments in the literature.

The clearest interpretation of the models in the statistical literature on dynamic treatment effects is as *ex post* selection-corrected analyses of distributions of events that have occurred. In a model of perfect certainty, where *ex post* and *ex ante* choices and outcomes are identical, the reduced-form approach can be interpreted as approximating clearly specified choice models. In a more general analysis with information arrival and agent updating of information sets, the nature of the approximation is less clear cut. Thus, the current reduced-form literature is unclear as to which agent decision-making processes and information arrival assumptions justify the conditional sequential randomization assumptions widely used in the dynamic treatment effect literature (see, e.g., Gill and Robins, 2001; Lechner and Miquel, 2002; Lok, 2007; Robins, 1989, 1997; Van der Laan and Robins, 2003). Section 24.3.2.2 provides some insight by highlighting the connection to the conditional-independence assumption often employed in the structural dynamic discrete-choice literature (see Rust, 1987; and the survey in Rust, 1994). Reduced-form approaches are not clear about the source of the unobservables and their relationship with conditioning variables. It would be a valuable exercise to exhibit which structural models are approximated by various reduced-form models. In the structural analysis, this specification emerges as part of the analysis, as our discussion of the stochastic properties of the unobservables presented in the preceding section makes clear.

The HN analysis of both structural and reduced-form models relies heavily on limit set arguments. They solve the selection problem in limit sets. The dynamic matching models of Gill and Robins (2001) and Lok (2007) solve the selection problem by invoking recursive conditional independence assumptions. In the context of the models of HN, they assume that the econometrician knows the θ or can eliminate the effect of θ on estimates of the model by conditioning on a suitable set of variables. The HN analysis entertains the possibility that analysts know less than the agents they study. It allows for some of the variables that would make matching valid to be unobservable. As we have noted in early subsections, versions of recursive conditional independence assumptions are also used in the dynamic discrete-choice literature (see the survey in Rust, 1994). The HN factor models allow analysts

to construct the joint distribution of outcomes across stopping times. This feature is missing from the statistical treatment effect literature.

Both HN's structural and reduced-form models of treatment choice are stopping time models. Neither model allows for multiple entry into and exit from treatment, even though agents in these models would like to reverse their treatment decisions for some realizations of their index if this was not too costly (or, in the case of the reduced-form model, if the index thresholds for returning would not be too low).[81] Cunha, Heckman, and Navarro (2007) derive conditions on structural stopping models from a more basic model that entertains the possibility of return from dropout states but which nonetheless exhibits the stopping time property. The HN identification strategy relies on the nonrecurrent nature of treatment. Their identification strategy of using limit sets can be applied to a recurrent model provided that analysts confine attention to subsets of (X,Z) such that in those subsets the probability of recurrence is zero.

24.5.5 A Short Survey of Dynamic Discrete-Choice Models

Rust (1994) presents a widely cited nonparametric nonidentification theorem for dynamic discrete-choice models. It is important to note the restrictive nature of his negative results. He analyzes a recurrent state infinite horizon model in a stationary environment. He does not use any exclusion restrictions or cross outcome-choice restrictions. He uses a general utility function. He places no restrictions on period-specific utility functions such as concavity or linearity nor does he specify restrictions connecting preferences and outcomes. One can break Rust's nonidentification result with additional information.

Magnac and Thesmar (2002) present an extended comment on Rust's analysis including positive results for identification when the econometrician knows the distributions of unobservables, assumes that unobservables enter period-specific utility functions in an additively separable way and is willing to specify functional forms of utility functions or other ingredients of the model, as do Pakes (1986), Keane and Wolpin (1997), Eckstein and Wolpin (1999) and Hotz and Miller (1988, 1993). Magnac and Thesmar (2002) also consider the case where one state (choice) is absorbing (as do Hotz and Miller, 1993) and where the value functions are known at the terminal age ($\bar{T}$) (as do Keane and Wolpin, 1997 and Belzil and Hansen, 2002). In HN, each treatment time is an absorbing state. In a separate analysis, Magnac and Thesmar consider the case where unobservables from the point of view of the econometrician are correlated over time (or age t) and choices (s) under the assumption that the distribution of the unobservables is known. They also consider the case where exclusion restrictions are available. Throughout their analysis, they maintain

[81] Recall that treatment occurs if the index turns positive. If there are costs to reversing this decision, agents would only reverse their decision if the index falls below some negative threshold. The stopping time assumption is equivalent to the assumption that the costs of reversal are prohibitively large, or that the corresponding threshold is at the lower end of the support of the index.

that the distribution of the unobservables is known both by the agent and the econometrician.

HN provide semiparametric identification of a finite-horizon finite-state model with an absorbing state with semiparametric specifications of reward and cost functions.[82] Given that rewards are in value units, the scale of their utility function is fixed as they also are in models of profit-maximizing firms. Choices are not invariant to arbitrary affine transformations so that one source of nonidentifiability in Rust's analysis is eliminated. They can identify the error distributions nonparametrically given their factor structure. They do not have to assume either the functional form of the unobservables or knowledge of the entire distribution of unobservables.

HN present a fully specified structural model of choices and outcomes motivated by, but not identical to, the analyses of Keane and Wolpin (1994, 1997) and Eckstein and Wolpin (1999). In their setups, outcome and cost functions are parametrically specified. Their states are recurrent while those of HN are absorbing. In the HN model, once an agent drops out of school, the agent does not return. In the Keane–Wolpin model, an agent who drops out can return. Keane and Wolpin do not establish identification of their model, whereas HN establish semiparametric identification of their model. They analyze models with more general times series processes for unobservables. In both the HN and Keane–Wolpin frameworks, agents learn about unobservables. In the Keane–Wolpin framework, such learning is about temporally independent shocks that do not affect agent expectations about returns relevant to possible future choices. The information just affects the opportunity costs of current choices. In the HN framework, learning affects agent expectations about future returns as well as current opportunity costs.

The HN model extends previous work by Carneiro et al. (2003); Cunha and Heckman (2007, 2008) and Cunha et al. (2005, 2006) by considering explicit multiperiod dynamic models with information updating. They consider one-shot decision models with information updating and associated outcomes.

Their analysis is related to that of Taber (2000). Like Cameron and Heckman (1998), both HN and Taber use identification-in-the-limit arguments.[83] Taber considers identification of a two period model with a general utility function, whereas in Sect. 24.5.2, we discuss how HN consider identification of a specific form of the utility function (an earnings function) for a multiperiod maximization problem. As in HN, Taber allows for the sequential arrival of information. His analysis is based on conventional exclusion restrictions, but the analysis of HN is not. They use outcome data in conjunction with the discrete dynamic choice data to exploit cross-equation restrictions, whereas Taber does not.

The HN treatment of serially correlated unobservables is more general than any discussion that appears in the current dynamic discrete choice and dynamic treatment effect literature. They do not invoke the strong sequential conditional independence assumptions used in the dynamic treatment effect literature in statistics (Gill

[82] Although their main theorems are for additively separable reward and cost functions, it appears that additive separability can be relaxed using the analysis of Matzkin (2003).

[83] Pakes and Simpson (1989) sketch a proof of identification of a model of the option values of patents that is based on limit sets for an option model.

and Robins, 2001; Lechner and Miquel, 2002; Lok, 2007; Robins, 1989, 1997), nor do they invoke the closely related conditional temporal independence of unobserved state variables given observed state variables invoked by Rust (1987); Hotz and Miller (1988, 1993); Manski (1993) and Magnac and Thesmar (2002) (in the first part of their paper) or the independence assumptions invoked by Wolpin (1984).[84] HN allow for more general time series dependence in the unobservables than is entertained by Pakes (1986), Keane and Wolpin (1997) or Eckstein and Wolpin (1999).[85]

Like Miller (1984) and Pakes (1986), HN explicitly model, identify and estimate agent learning that affects expected future returns.[86] Pakes and Miller assume functional forms for the distributions of the error process and for the serial correlation pattern about information updating and time series dependence. The HN analysis of the unobservables is nonparametric and they estimate, rather than impose, the stochastic structure of the information updating process.

Virtually all papers in the literature, including the HN analysis, invoke rational expectations. An exception is the analysis of Manski (1993) who replaces rational expectations with a synthetic cohort assumption that choices and outcomes of one group can be observed (and acted on) by a younger group. This assumption is more plausible in stationary environments and excludes any temporal dependence in unobservables. In recent work, Manski (2004) advocates use of elicited expectations as an alternative to the synthetic cohort approach.

While HN use rational expectations, they estimate, rather than impose the structure of agent information sets. Miller (1984), Pakes (1986), Keane and Wolpin (1997), and Eckstein and Wolpin (1999) assume that they know the law governing the evolution of agent information up to unknown parameters.[87] Following the procedure presented in Cunha and Heckman (2007, 2008); Cunha et al. (2005, 2006) and Navarro (2005), HN can test for which factors (θ) appear in agent information sets at different stages of the life cycle and they identify the distributions of the unobservables nonparametrically.

The HN analysis of dynamic treatment effects is comparable, in some aspects, to the recent continuous-time event-history approach of Abbring and Van den Berg (2003b) previously analyzed. Those authors build a continuous-time model of counterfactuals for outcomes that are durations. They model treatment assignment times using a continuous-time duration model.

[84] Manski (1993) and Hotz and Miller (1993) use a synthetic cohort effect approach that assumes that young agents will follow the transitions of contemporaneous older agents in making their lifecycle decisions. Manski and Hotz and Miller exclude any temporally dependent unobservables from their models. The synthetic cohort approach has been widely used in labor economics at least since Mincer (1974). See Ghez and Becker (1975), MaCurdy (1981) and Mincer (1974) for applications of the synthetic cohort approach. For empirical evidence against the assumption that the earnings of older workers are a reliable guide to the earnings of younger workers in models of earnings and schooling choices for recent cohorts of workers, see Heckman et al. (2006).

[85] Rust (1994) provides a clear statement of the stochastic assumptions underlying the dynamic discrete-choice literature up to the date of his survey.

[86] As previously noted, the previous literature assumes learning only about current costs.

[87] They specify a priori particular processes of information arrival as well as which components of the unobservables agents know and act on, and which components they do not.

The HN analysis is in discrete time and builds on previous work by Heckman (1981a,c) on heterogeneity and state dependence that identifies the causal effect of employment (or unemployment) on future employment (or unemployment).[88] They model time to treatment and associated vectors of outcome equations that may be discrete, continuous or mixed discrete-continuous. In a discrete-time setting, they are able to generate a variety of distributions of counterfactuals and economically motivated parameters. They allow for heterogeneity in responses to treatment that has a general time series structure.

As noted in Sect. 24.5.4, Abbring and Van den Berg (2003b) do not identify explicit agent information sets as HN do in their paper and as is done in Cunha et al. (2005), and they do not model learning about future rewards. Their outcomes are restricted to be continuous-time durations. The HN framework is formulated in discrete time, which facilitates the specification of richer unobserved and observed covariate processes than those entertained in the continuous-time framework of Abbring and Van den Berg (2003b). It is straightforward to attach a vector of treatment outcomes in the HN model that includes continuous outcomes, discrete outcomes and durations expressed as binary strings.[89] At a practical level, the approach often can produce very fine-grained descriptions of continuous-time phenomena by using models with many finite periods. Clearly, a synthesis of the event-history approach with the HN approach would be highly desirable. That would entail taking continuous-time limits of the discrete-time models. It is a task that awaits completion.

Flinn and Heckman (1982) utilize information on stopping times and associated wages to derive cross-equation restrictions to partially identify an equilibrium job search model for a stationary economic environment where agents have an infinite horizon. They establish that the model is nonparametrically nonidentified. Their analysis shows that use of outcome data in conjunction with data on stopping times is not sufficient to secure nonparametric identification of a dynamic discrete-choice model, even when the reward function is linear in outcomes unlike the reward functions in Rust (1987) and Magnac and Thesmar (2002). Parametric restrictions can break their nonidentification result. Abbring and Campbell (2005) exploit such restrictions, together with cross-equation restrictions on stopping times and noisy outcome measures, to prove identification of an infinite-horizon model of firm survival and growth with entrepreneurial learning. Alternatively, nonstationarity arising from finite horizons can break their nonidentification result (see Wolpin, 1987). The HN analysis exploits the finite-horizon backward-induction structure of our model in conjunction with outcome data to secure identification and does not rely on arbitrary period by period exclusion restrictions. They substantially depart from the assumptions maintained in Rust's nonidentification theorem (1994). They achieve identification by using cross-equation restrictions, linearity of preferences

[88] Heckman and Borjas (1980) investigate these issues in a continuous-time duration model. See also Heckman and MaCurdy (1980).

[89] Abbring (2008) considers nonparametric identification of mixed semi-Markov event-history models that extends his work with Van den Berg. See Sect. 24.4.

and additional measurements, and exploiting the structure of their finite horizon nonrecurrent model. Nonstationarity of regressors greatly facilitates identification by producing both exclusion and curvature restrictions which can substitute for standard exclusion restrictions.

24.6 Conclusion

This paper has surveyed recent approaches to using panel data to evaluate policies. We have compared and contrasted the statistical dynamic treatment approach based on sequential conditional independence assumptions that generalize matching to approaches developed in econometrics. We compared and contrasted a continuous-time event-history approach developed by Abbring and Van den Berg (2003b) to discrete-time reduced-form and structural models developed by Heckman and Navarro (2007), and Cunha et al. (2005).

References

Aakvik, A., J. J. Heckman, and E. J. Vytlacil (2005). Estimating treatment effects for discrete outcomes when responses to treatment vary: An application to Norwegian vocational rehabilitation programs. *Journal of Econometrics 125*(1–2), 15–51.

Aalen, O. O. and H. K. Gjessing (2004, December). Survival models based on the Ornstein-Uhlenbeck process. *Lifetime Data Analysis 10*(4), 407–423.

Abbring, J. H. (2002, February). Stayers versus defecting movers: A note on the identification of defective duration models. *Economics Letters 74*(3), 327–331.

Abbring, J. H. (2003). Dynamic econometric program evaluation. Discussion Paper 804, IZA, Bonn. Paper prepared for the H. Theil Memorial Conference, Amsterdam, 16–18 August 2002.

Abbring, J. H. (2007). Mixed hitting-time models. Discussion Paper 07-57/3, Tinbergen Institute, Amsterdam.

Abbring, J. H. (2008). The event-history approach to program evaluation. In D. Millimet, J. Smith, and E. Vytlacil (Eds.), *Modeling and Evaluating Treatment Effects in Econometrics*, Volume 21 of *Advances in Econometrics*. Oxford: Elsevier Science, pp. 33–55.

Abbring, J. H. and J. R. Campbell (2005). A firm's first year. Discussion Paper 05-046/3, Tinbergen Institute, Amsterdam.

Abbring, J. H. and J. J. Heckman (2007). Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6, pp. 5145–5303, Amsterdam: Elsevier.

Abbring, J. H. and G. J. Van den Berg (2003a, September). The identifiability of the mixed proportional hazards competing risks model. *Journal of the Royal Statistical Association, Series B 65*(3), 701–710.

Abbring, J. H. and G. J. Van den Berg (2003b, September). The nonparametric identification of treatment effects in duration models. *Econometrica 71*(5), 1491–1517.

Abbring, J. H. and G. J. Van den Berg (2003c). A simple procedure for inference on treatment effects in duration models. Discussion Paper 2003:19, IFAU, Uppsala.

Abbring, J. H. and G. J. Van den Berg (2004, January). Analyzing the effect of dynamically assigned treatments using duration models, binary treatment models, and panel data models. *Empirical Economics 29*(1), 5–20.

Abbring, J. H. and G. J. Van den Berg (2005). Social experiments and instrumental variables with duration outcomes. Discussion Paper 05-047/3, Tinbergen Institute, Amsterdam.

Abbring, J. H., G. J. Van den Berg, and J. C. Van Ours (2005, July). The effect of unemployment insurance sanctions on the transition rate from unemployment to employment. *Economic Journal 115*(505), 602–630.

Aldrich, J. (1989, January). Autonomy. *Oxford Economic Papers 41*(1), 15–34.

Andersen, P. K., Ø. Borgan, R. Gill, and N. Keiding (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Belzil, C. and J. Hansen (2002, September). Unobserved ability and the return to schooling. *Econometrica 70*(5), 2075–2091.

Black, D. A., J. A. Smith, M. C. Berger, and B. J. Noel (2003, September). Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system. *American Economic Review 93*(4), 1313–1327.

Bonnal, L., D. Fougère, and A. Sérandon (1997, October). Evaluating the impact of French employment policies on individual labour market histories. *Review of Economic Studies 64*(4), 683–713.

Cameron, S. V. and J. J. Heckman (1998, April). Life cycle schooling and dynamic selection bias: Models and evidence for five cohorts of American males. *Journal of Political Economy 106*(2), 262–333.

Card, D. and D. G. Sullivan (1988, May). Measuring the effect of subsidized training programs on movements in and out of employment. *Econometrica 56*(3), 497–530.

Carneiro, P., K. Hansen, and J. J. Heckman (2001, Fall). Removing the veil of ignorance in assessing the distributional impacts of social policies. *Swedish Economic Policy Review 8*(2), 273–301.

Carneiro, P., K. Hansen, and J. J. Heckman (2003, May). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review 44*(2), 361–422.

Chan, T. Y. and B. H. Hamilton (2006). Learning, private information and the economic evaluation of randomized experiments. *Journal of Political Economy 114*(6), 997–1040.

Cunha, F. and J. J. Heckman (2007). The evolution of earnings risk in the US economy. Presented at the 9th World Congress of the Econometric Society, London 2005, under revision.

Cunha, F. and J. J. Heckman (2008). A framework for the analysis of inequality. *Macroeconomic Dynamics*. Forthcoming.

Cunha, F., J. J. Heckman, and S. Navarro (2005, April). Separating uncertainty from heterogeneity in life cycle earnings, The 2004 Hicks Lecture. *Oxford Economic Papers 57*(2), 191–261.

Cunha, F., J. J. Heckman, and S. Navarro (2006). Counterfactual analysis of inequality and social mobility. In S. L. Morgan, D. B. Grusky, and G. S. Fields (Eds.), *Mobility and Inequality: Frontiers of Research in Sociology and Economics*, Chap. 4, pp. 290–348. Stanford, CA: Stanford University Press.

Cunha, F., J. J. Heckman, and S. Navarro (2007). The identification and economic content of ordered choice models with stochastic cutoffs. *International Economic Review*. *48*(4), 1273–1309.

Eberwein, C., J. C. Ham, and R. J. LaLonde (1997, October). The impact of being offered and receiving classroom training on the employment histories of disadvantaged women: Evidence from experimental data. *Review of Economic Studies 64*(4), 655–682.

Eckstein, Z. and K. I. Wolpin (1999, November). Why youths drop out of high school: The impact of preferences, opportunities, and abilities. *Econometrica 67*(6), 1295–1339.

Falmagne, J.-C. (1985). *Elements of Psychophysical Theory*. Oxford Psychology Series No. 6. New York: Oxford University Press.

Fitzenberger, B., A. Osikominu, and R. Völter (2006, May). Get training or wait? Long-run employment effects of training programs for the unemployed in West Germany, Discussion paper 2121, IZA, Bonn.

Fleming, T. R. and D. P. Harrington (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

Flinn, C. and J. J. Heckman (1982, January). New methods for analyzing structural models of labor force dynamics. *Journal of Econometrics 18*(1), 115–68.

Florens, J.-P. and M. Mouchart (1982). A note on noncausality. *Econometrica 50*, 583–591.

Freedman, D. A. (2004, August). On specifying graphical models for causation and the identification problem. *Evaluation Review 28*(4), 267–293.

Freund, J. E. (1961, December). A bivariate extension of the exponential distribution. *Journal of the American Statistical Association 56*(296), 971–977.

Frisch, R. (1938). Autonomy of economic relations. Paper given at League of Nations. Reprinted in D.F. Hendry and M.S. Morgan (1995), *The Foundations of Econometric Analysis*, Cambridge: Cambridge University Press.

Ghez, G. R. and G. S. Becker (1975). *The Allocation of Time and Goods over the Life Cycle*. New York: National Bureau of Economic Research.

Gill, R. D. and J. M. Robins (2001, December). Causal inference for complex longitudinal data: The continuous case. *Annals of Statistics 29*(6), 1785–1811.

Granger, C. W. J. (1969, August). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica 37*(3), 424–438.

Gritz, R. M. (1993, May–June). The impact of training on the frequency and duration of employment. *Journal of Econometrics 57*(1–3), 21–51.

Grubb, D. (2000). Eligibility criteria for unemployment benefits. In *OECD Economic Studies, Special issue: Making Work Pay*, Number 31, pp. 147–184. OECD.

Haavelmo, T. (1943, January). The statistical implications of a system of simultaneous equations. *Econometrica 11*(1), 1–12.

Ham, J. C. and R. J. LaLonde (1996, January). The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training. *Econometrica 64*(1), 175–205.

Hansen, L. P. and T. J. Sargent (1980, February). Formulating and estimating dynamic linear rational expectations models. *Journal of Economic Dynamics and Control 2*(1), 7–46.

Heckman, J. J. (1974, July). Shadow prices, market wages, and labor supply. *Econometrica 42*(4), 679–694.

Heckman, J. J. (1981a). Heterogeneity and state dependence. In S. Rosen (Ed.), *Studies in Labor Markets*, National Bureau of Economic Research, pp. 91–139. Chicago: University of Chicago Press.

Heckman, J. J. (1981b). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process and some Monte Carlo evidence. In C. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 179–85. Cambridge, MA: MIT Press.

Heckman, J. J. (1981c). Statistical models for discrete panel data. In C. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 114–178. Cambridge, MA: MIT Press.

Heckman, J. J. (1990, May). Varieties of selection bias. *American Economic Review 80*(2), 313–318.

Heckman, J. J. (2005a). The scientific model of causality. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J. (2005b, August). The scientific model of causality. *Sociological Methodology 35*(1), 1–97.

Heckman, J. J. and G. J. Borjas (1980, August). Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence. *Economica 47*(187), 247–283. Special Issue on Unemployment.

Heckman, J. J. and B. E. Honoré (1989, June). The identifiability of the competing risks model. *Biometrika 76*(2), 325–330.

Heckman, J. J. and B. E. Honoré (1990, September). The empirical content of the Roy model. *Econometrica 58*(5), 1121–1149.

Heckman, J. J., R. J. LaLonde, and J. A. Smith (1999). The economics and econometrics of active labor market programs. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3A, Chap. 31, pp. 1865–2097. New York: North-Holland.

Heckman, J. J., L. J. Lochner, and C. Taber (1998, January). Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. *Review of Economic Dynamics 1*(1), 1–58.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2006). Earnings equations and rates of return: The Mincer equation and beyond. In E. A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, Chap. 7, pp. 307–458. Amsterdam: North-Holland.

Heckman, J. J. and T. E. MaCurdy (1980, January). A life cycle model of female labour supply. *Review of Economic Studies 47*(1), 47–74.

Heckman, J. J. and S. Navarro (2004, February). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics 86*(1), 30–57.

Heckman, J. J. and S. Navarro (2005). Empirical estimates of option values of education and information sets in a dynamic sequential choice model. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J. and S. Navarro (2007, February). Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics 136*(2), 341–396.

Heckman, J. J. and B. S. Singer (1984, January–February). Econometric duration analysis. *Journal of Econometrics 24*(1–2), 63–132.

Heckman, J. J. and B. S. Singer (1986). Econometric analysis of longitudinal data. In Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics*, Volume 3, Chap. 29, pp. 1690–1763. Amsterdam: North-Holland.

Heckman, J. J. and J. A. Smith (1998). Evaluating the welfare state. In S. Strom (Ed.), *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, pp. 241–318. New York: Cambridge University Press.

Heckman, J. J., J. Stixrud, and S. Urzua (2006, July). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics 24*(3), 411–482.

Heckman, J. J. and C. Taber (1994). Econometric mixture models and more general models for unobservables in duration analysis. *Statistical Methods in Medical Research 3*(3), 279–299.

Heckman, J. J., J. L. Tobias, and E. J. Vytlacil (2001, October). Four parameters of interest in the evaluation of social programs. *Southern Economic Journal 68*(2), 210–223.

Heckman, J. J., J. L. Tobias, and E. J. Vytlacil (2003, August). Simple estimators for treatment parameters in a latent variable framework. *Review of Economics and Statistics 85*(3), 748–754.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics 88*(3), 389–432.

Heckman, J. J., S. Urzua, and G. Yates (2007). The identification and estimation of option values in a model with recurrent states. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J. and E. J. Vytlacil (1999, April). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences 96*, 4730–4734.

Heckman, J. J. and E. J. Vytlacil (2001). Causal parameters, treatment effects and randomization. Unpublished manuscript, University of Chicago, Department of Economics.

Heckman, J. J. and E. J. Vytlacil (2005, May). Structural equations, treatment effects and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (2007a). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 4779–4874, Amsterdam: Elsevier.

Heckman, J. J. and E. J. Vytlacil (2007b). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 4875–5144, Amsterdam: Elsevier.

Hendry, D. F. and M. S. Morgan (1995). *The Foundations of Econometric Analysis*. New York: Cambridge University Press.

Holland, P. W. (1986, December). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Honoré, B. E. (1993, January). Identification results for duration models with multiple spells. *Review of Economic Studies 60*(1), 241–246.

Honoré, B. E. and A. Lewbel (2002, September). Semiparametric binary choice panel data models without strictly exogenous regressors. *Econometrica 70,*(5), 2053–2063.

Hotz, V. J. and R. A. Miller (1988, January). An empirical analysis of life cycle fertility and female labor supply. *Econometrica 56*(1), 91–118.

Hotz, V. J. and R. A. Miller (1993, July). Conditional choice probabilities and the estimation of dynamic models. *Review of Economic Studies 60*(3), 497–529.

Hurwicz, L. (1962). On the structural form of interdependent systems. In E. Nagel, P. Suppes, and A. Tarski (Eds.), *Logic, Methodology and Philosophy of Science*, pp. 232–239. Palo Alto: Stanford University Press.

Kalbfleisch, J. D. and R. L. Prentice (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Keane, M. P. and K. I. Wolpin (1994, November). The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence. *The Review of Economics and Statistics 76*(4), 648–672.

Keane, M. P. and K. I. Wolpin (1997, June). The career decisions of young men. *Journal of Political Economy 105*(3), 473–522.

Keiding, N. (1999, September). Event history analysis and inference from observational epidemiology. *Statistics in Medicine 18*(17–18), 2353–2363.

Lancaster, T. (1979, July). Econometric methods for the duration of unemployment. *Econometrica 47*(4), 939–956.

Leamer, E. E. (1985, Spring). Vector autoregressions for causal inference? *Carnegie-Rochester Conference Series on Public Policy 22*, 255–303.

Lechner, M. and R. Miquel (2002). Identification of effects of dynamic treatments by sequential conditional independence assumptions. Discussion paper, University of St. Gallen, Department of Economics.

Lewbel, A. (2000, July). Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics 97*(1), 145–177.

Lok, J. (2007). Statistical modelling of causal effects in continuous time. *Annals of Statistics*. Forthcoming.

MaCurdy, T. E. (1981, December). An empirical model of labor supply in a life-cycle setting. *Journal of Political Economy 89*(6), 1059–1085.

Magnac, T. and D. Thesmar (2002, March). Identifying dynamic discrete decision processes. *Econometrica 70*(2), 801–816.

Manski, C. F. (1988, September). Identification of binary response models. *Journal of the American Statistical Association 83*(403), 729–738.

Manski, C. F. (1993, July). Dynamic choice in social settings: Learning from the experiences of others. *Journal of Econometrics 58*(1–2), 121–136.

Manski, C. F. (2004, September). Measuring expectations. *Econometrica 72*(5), 1329–1376.

Matzkin, R. L. (1992, March). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica 60*(2), 239–270.

Matzkin, R. L. (1993, July). Nonparametric identification and estimation of polychotomous choice models. *Journal of Econometrics 58*(1–2), 137–168.

Matzkin, R. L. (1994). Restrictions of economic theory in nonparametric methods. In R. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, pp. 2523–58. New York: North-Holland.

Matzkin, R. L. (2003, September). Nonparametric estimation of nonadditive random functions. *Econometrica 71*(5), 1339–1375.

Meyer, B. D. (1990, July). Unemployment insurance and unemployment spells. *Econometrica 58*(4), 757–782.

Meyer, B. D. (1996, January). What have we learned from the Illinois reemployment bonus experiment? *Journal of Labor Economics 14*(1), 26–51.

Miller, R. A. (1984, December). Job matching and occupational choice. *Journal of Political Economy 92*(6), 1086–1120.

Mincer, J. (1974). *Schooling, Experience and Earnings*. New York: Columbia University Press for National Bureau of Economic Research.

Mortensen, D. T. (1977, July). Unemployment insurance and job search decisions. *Industrial and Labor Relations Review 30*(4), 505–517.

Mortensen, D. T. (1986). Job search and labor market analysis. In O. Ashenfelter and P. R. C. Layard (Eds.), *Handbook of Labor Economics*, Volume 2 of *Handbooks in Economics*, pp. 849–919. New York: Elsevier Science.

Murphy, S. A. (2003, May). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Association, Series B 65*(2), 331–366.

Navarro, S. (2004). Semiparametric identification of factor models for counterfactual analysis. Unpublished manuscript, University of Chicago, Department of Economics.

Navarro, S. (2005). *Understanding Schooling: Using Observed Choices to Infer Agent's Information in a Dynamic Model of Schooling Choice When Consumption Allocation is Subject to Borrowing Constraints*. Ph.D. Dissertation, University of Chicago, Chicago, IL.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. *Roczniki Nauk Rolniczych 10*, 1–51. in Polish; edited and translated version of Sect. 9 by D.M. Dabrowska and T.P Speed (1990), *Statistical Science*, **5**, 465–472.

Pakes, A. (1986, July). Patents as options: Some estimates of the value of holding european patent stocks. *Econometrica 54*(4), 755–784.

Pakes, A. and M. Simpson (1989). Patent renewal data. *Brookings Papers on Economic Activity* (Special Issue), 331–401.

Pearl, J. (2000). *Causality*. Cambridge, England: Cambridge University Press.

Quandt, R. E. (1958, December). 'The estimation of the parameters of a linear regression system obeying two separate regimes'. *Journal of the American Statistical Association 53*(284), 873–880.

Quandt, R. E. (1972, June). 'A new approach to estimating switching regressions'. *Journal of the American Statistical Association 67*(338), 306–310.

Ridder, G. (1986, April). An event history approach to the evaluation of training, recruitment and employment programmes. *Journal of Applied Econometrics 1*(2), 109–126.

Ridder, G. (1990, April). The non-parametric identification of generalized accelerated failure-time models. *Review of Economic Studies 57*(2), 167–181.

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. In L. Sechrest, H. Freeman, and A. Mulley (Eds.), *Health Services Research Methodology: A Focus on AIDS*, pp. 113–159. Rockville, MD: U.S. Department of Health and Human Services, National Center for Health Services Research and Health Care Technology Assessment.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics*, pp. 69–117. New York: Springer-Verlag.

Rosenzweig, M. R. and K. I. Wolpin (2000, December). Natural "natural experiments" in economics. *Journal of Economic Literature 38*(4), 827–874.

Roy, A. (1951, June). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*(2), 135–146.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association 81*(396), 961–962.

Rust, J. (1987, September). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica 55*(5), 999–1033.

Rust, J. (1994). Structural estimation of Markov decision processes. In R. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, pp. 3081–3143. New York: North-Holland.

Sims, C. A. (1972, September). Money, income, and causality. *American Economic Review 62*(4), 540–552.

Taber, C. R. (2000, June). Semiparametric identification and heterogeneity in discrete choice dynamic programming models. *Journal of Econometrics 96*(2), 201–229.

Thurstone, L. L. (1930). *The Fundamentals of Statistics*. New York: The Macmillan Company.

Urzua, S. (2005). Schooling choice and the anticipation of labor market conditions: A dynamic choice model with heterogeneous agents and learning. Unpublished manuscript, University of Chicago, Department of Economics.

Van den Berg, G. J. (2001). Duration models: Specification, identification and multiple durations. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5 of *Handbooks in Economics*, pp. 3381–3460. New York: North-Holland.

Van den Berg, G. J., A. Holm, and J. C. Van Ours (2002, November). Do stepping-stone jobs exist? Early career paths in the medical profession. *Journal of Population Economics 15*(4), 647–665.

Van den Berg, G. J., B. Van der Klaauw, and J. C. Van Ours (2004, January). Punitive sanctions and the transition rate from welfare to work. *Journal of Labor Economics 22*(1), 211–241.

Van der Laan, M. J. and J. M. Robins (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer-Verlag.

Wolpin, K. I. (1984, October). An estimable dynamic stochastic model of fertility and child mortality. *Journal of Political Economy 92*(5), 852–874.

Wolpin, K. I. (1987, July). Estimating a structural search model: The transition from school to work. *Econometrica 55*(4), 801–817.

Chapter 25
Econometrics of Individual Labor Market Transitions

Denis Fougère and Thierry Kamionka

25.1 Introduction

During the last 20 years, the microeconometric analysis of individual transitions has been extensively used for investigating some problems inherent in the functioning of contemporary labor markets, such as the relations between individual mobility and wages, the variability of flows between employment, unemployment and non-employment through the business cycle, or the effects of public policies (training programs, unemployment insurance, ...) on individual patterns of unemployment. Typically, labor market transition data register sequences of durations spent by workers in distinct states, such as employment, unemployment and non-employment. When individual participation histories are completely observed through panel or retrospective surveys, the econometrician then disposes of continuous-time realizations of the labor market participation process. When these histories are only observed at many successive dates through panel surveys, the available information is a truncated one; more precisely it takes the form of discrete-time observations of underlying continuous-time processes. Our presentation of statistical procedures used for analysing individual transition or mobility histories is based on the distinction between these two kinds of data.

Denis Fougère
CNRS, CREST-INSEE (Paris), CEPR (London) and IZA (Bonn), e-mail: fougere@ensae.fr

Thierry Kamionka
CNRS and CREST-INSEE, Paris, CREST-INSEE 15, Boulevard Gabriel Péri, 92245 Malakoff cedex, France, e-mail: kamionka@ensae.fr

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

Statistical models of labor market transitions can be viewed as extensions of the single-spell unemployment duration model (see Chap. 14, this volume). Theoretically, a transition process is a continuous-time process taking its values in a finite discrete state space whose elements represent the main labor force participation states, for example employment, unemployment and non-employment.

The goal is then to estimate parameters which capture effects of different time-independent or time-varying exogenous variables on intensities of transition between states of participation. Here transition intensities represent conditional instantaneous probabilities of transition between two distinct states at some date. Typically, the analyst is interested in knowing the sign and the size of the influence of a given variable, such as the unemployment insurance amount or the past training and employment experiences, on the transition from unemployment to employment for example, and more generally in predicting the effect of such variables on the future of the transition process. For this purpose, she can treat these variables as regressors in the specification of transition intensities. Doing that, she estimates a reduced-form model of transition. Estimation of a more structural model requires the specification of an underlying dynamic structure in which the participation state is basically the choice set for a worker and in which parameters to be estimated influence directly individual objective functions (such as intertemporal utility functions) which must be maximized under some revelant constraints inside a dynamic programming setup. Such structural models have been surveyed by Eckstein and Wolpin (1989) or Rust (1994).

Our survey focuses only on reduced-form transition models, which have been extensively used and estimated in labor microeconometrics. The first section contains a general presentation of the statistical modelling of the transition process for continuous-time (event-history) data. The first section briefly recalls the useful mathematical definitions, essentially the ones characterizing the distribution of the joint sequence of visited states and of sojourn durations in these states. It also presents parametric and nonparametric estimation procedures, and ends with the issue of the unobserved heterogeneity treatment in this kind of process.

The second section deals with inference for a particular class of transition processes, namely markovian processes or simple mixtures of markovian processes, using discrete-time panel observations. Here the main problem is the embeddability of the discrete-time Markov chain into a continuous time one. In other words, the question is whether or not the discrete-time panel observations of a transition process are generated by a continuous-time homogeneous Markov process. After a discussion of this problem, the second section presents maximum-likelihood and bayesian procedures for estimating the transition intensity matrix governing the evolution of the continuous-time markovian process. Particular attention is paid to the estimation of the continuous-time mover–stayer model, which is the more elementary model of mixed Markov processes.

The conclusion points out some extensions.

25.2 Multi-spell Multi-state Models

25.2.1 General framework

25.2.1.1 Notations

Let us consider a cadlag[1] stochastic process X_t, $t \in \mathbb{R}^+$, taking its value in a finite discrete-state space denoted $E = \{1, \ldots, K\}$, $K \in \mathbb{N}$ and $K \geq 2$. In other words, K represents the total number of states for the process, and X_t is the state occupied at time t by the individual (so $X_t \in E, \forall t \in \mathbb{R}^+$). Let $\{x_t, t \in \mathbb{R}^+\}$ be a realization of this process. We suppose that all the individual realizations of this process are identically and independently distributed: to simplify the notations, we can then omit the index for individuals.

As an illustration we consider the case of a labor force participation process describing the state occupied by a worker at time t. In order to simplify, we set:

$$X_t = \begin{cases} 1 & \text{if the individual is employed at time } t \\ 2 & \text{if the individual is unemployed at time } t \\ 3 & \text{if the individual is out of the labor force at time } t \end{cases} \tag{25.1}$$

Now we suppose that each individual process is observed from the date of entry into the labor market, denoted τ_0 for the individual, up to an exogenously fixed time τ_e ($\tau_e > \tau_0$). An example of realization of process X_t is represented in Fig. 25.1.

This figure shows that the individual is first employed from time τ_0 up to time τ_1, then unemployed from time τ_1 up to time τ_2, then employed once again from

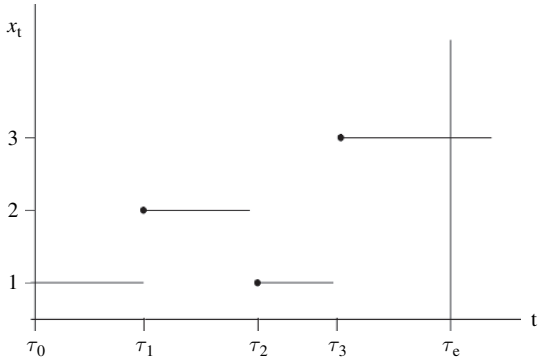

Fig. 25.1 A realization of the process X_t

[1] "cadlag" means right-continuous, admitting left limits. For the definition of a cadlag process, see Chap. 17, Sect. 17.1, this volume.

time τ_2 up to time τ_3, and finally out of the labor force (from time τ_3 on) when the observation stops at time τ_e. If we denote:

$$u_\ell = \tau_\ell - \tau_{\ell-1} \ , \ \ell = 1,2,\ldots \tag{25.2}$$

the sojourn duration in state $x_{\tau_{(\ell-1)}}$ reached by the individual at time $\tau_{(\ell-1)}$ (before a transition to state x_{τ_ℓ} at time τ_ℓ), the process X_t can be equivalently characterized by the sequences $\left\{ (\tau_\ell,\, x_{\tau_\ell}) ; \ell \in \mathbb{N} \right\}$ or $\{(u_\ell,\, x_{\sum_{k=0}^{\ell} u_k})_\ell \ ; \ \ell \in \mathbb{N}\}$ with $u_0 = \tau_0$.

Now suppose that process X_t is observed from the exogenous date τ_s, with $\tau_s \in]\tau_0,\ \tau_1[$, up to time τ_e and that the date of entry into the state occupied at time τ_s (i.e. the date of entry into the labor market, τ_0) is unknown to the analyst. Then, the sojourn duration in state $x_{\tau_s} = x_{\tau_0}$ is said to be *left-censored*. Symmetrically, for the example in Fig. 25.1, the sojourn duration in state $x_{\tau_e} = x_{\tau_3}$ is said to be *right-censored*, because the couple $(\tau_4,\, x_{\tau_4})$ is not observed.

We restrict now our attention to non left-censored samples, i.e. such that $\tau_s = \tau_0$, for all individuals.[2] We define the *event-history* corresponding to process X_t for the observation period $[\tau_0,\ \tau_e]$ as:

$$\omega \ = \ \left\{ \tau_0,\, x_{\tau_0},\, \tau_1,\, x_{\tau_1}, \ldots,\, \tau_n,\, x_{\tau_n} \right\} \tag{25.3}$$

where n is the number of transitions, i.e. the number of modifications, of the studied process during the period $[\tau_0,\ \tau_e]$. This event-history can be equivalently defined as:

$$\omega = \left\{ \tau_0, u_1, x_{\tau_0+u_1}, u_2, x_{\tau_0+u_1+u_2}, \ldots, u_n, x_{\tau_0+\sum_{\ell=1}^{n} u_\ell} \right\} \tag{25.4}$$

This realization of the process from time τ_0 to time τ_e can be written:

$$\omega = ((\tau_0, x_{\tau_0}), (u_1, x_{\tau_1}), \ldots, (u_n, x_{\tau_n}), (u_{n+1}, 0)) \tag{25.5}$$

where $u_{n+1} = \tau_e - \tau_n$ is the duration of the last observed spell. The last spell is right-censored. Indeed, τ_{n+1} and x_{n+1} are not observed. Consequently, we fix $x_{n+1} = 0$ in order to signify that the last duration is at least equal to u_{n+1}. This realization of the process can be rewritten

$$\omega = (y_0, y_1, \ldots, y_n, y_{n+1}) \tag{25.6}$$

where

$$y_k = \begin{cases} (\tau_0, x_{\tau_0}) & \text{if } \ k = 0 \\ (\tau_k, x_{\tau_k}) & \text{if } \ 1 \leq k \leq n \\ (\tau_{n+1}, 0) & \text{if } \ k = n+1 \end{cases}$$

Let us define a spell as a period of time delimited by two successive transitions. The history of the process is a sequence of variables $y_k = (u_k, x_{\tau_k})$, where u_k is the length of spell k and x_{τ_k} is the state occupied by the individual at time τ_k.

[2] The statistical treatment of left-censored spells has been considered by Heckman and Singer (1984), Ondrich 1985 and Amemiya 2001.

25.2.1.2 Distributions of Spell Durations

Suppose now that the process enters state $x_{\tau_{\ell-1}}$ ($x_{\tau_{\ell-1}} \in \{1, \ldots, K\}$) at time $\tau_{\ell-1}$ ($\ell = 1, \ldots, n+1$). Let us examine the probability distribution of the sojourn duration in state $x_{\tau_{\ell-1}}$ entered after the $(\ell-1)$-th transition of the process. For that purpose, we assume that this sojourn duration is generated by a conditional probability distribution P given the event-history $(y_0, \ldots, y_{\ell-1})$ and a vector of exogenous variables z, defined by the cumulative distribution function

$$\begin{aligned} F(u \mid y_0, \ldots, y_{\ell-1}; z; \theta) &= \Pr[U_\ell \leq u \mid y_0, \ldots, y_{\ell-1}; z; \theta] \\ &= 1 - S(u \mid y_0, \ldots, y_{\ell-1}; z; \theta) \end{aligned} \tag{25.7}$$

where θ is a vector of unknown parameters. Here U_ℓ denotes the random variable corresponding to the duration of the $\ell - th$ spell of the process, starting with its $(\ell-1) - th$ transition. $S(u \mid y_0, \ldots, y_{\ell-1}; z; \theta)$ is the *survivor function* of the sojourn duration in the $\ell - th$ spell. If the probability distribution P admits a density f with respect to the Lebesgue measure, then:

$$F(u \mid y_0, \ldots, y_{\ell-1}; z; \theta) = \int_0^u f(t \mid y_0, \ldots, y_{\ell-1}; z; \theta)\, \mathrm{d}t \tag{25.8}$$

and

$$\begin{aligned} f(u \mid y_0, \ldots, y_{\ell-1}; z; \theta) &= \frac{d}{\mathrm{d}u} F(u \mid y_0, \ldots, y_{\ell-1}; z; \theta) \\ &= -\frac{d}{\mathrm{d}u} S(u \mid y_0, \ldots, y_{\ell-1}; z; \theta) \end{aligned} \tag{25.9}$$

If the function $f(u \mid y_0, \ldots, y_{\ell-1}; z; \theta)$ is cadlag, then there exists a function, called the *hazard function* of the sojourn duration in the $\ell - th$ spell, defined as

$$\begin{aligned} h(u \mid y_0, \ldots, y_{\ell-1}; z; \theta) &= \frac{f(u \mid y_0, \ldots, y_{\ell-1}; z; \theta)}{S(u \mid y_0, \ldots, y_{\ell-1}; z; \theta)} \\ &= -\frac{d}{du} \log\, S(u \mid y_0, \ldots, y_{\ell-1}; z; \theta) \end{aligned} \tag{25.10}$$

or equivalently as

$$h(u|y_0, \ldots, y_{\ell-1}; z; \theta)\, du = \lim_{d\,u \downarrow 0} \frac{\Pr[u \leq U_\ell < u + d\,u \mid U_\ell \geq u; y_0, \ldots, y_{\ell-1}; z]}{d\,u} \tag{25.11}$$

From (25.9), it follows that:

$$\begin{aligned} -\log\, S(u \mid\ y_0, \ldots, y_{\ell-1}; z; \theta) &= \int_0^u h(t \mid y_0, \ldots, y_{\ell-1}; z; \theta)\, \mathrm{d}t \\ &= H(u \mid y_0, \ldots, y_{\ell-1}; z; \theta) \end{aligned} \tag{25.12}$$

The function $H_\ell(u \mid y_0,\ldots,y_{\ell-1};z)$ is called the conditional *integrated hazard function* of the sojourn in the $\ell-th$ spell, given the history of the process up to time $\tau_{\ell-1}$.

Reduced-form statistical models of labour-market transitions can be viewed as extensions of competing risks duration models or multi-states multi-spells duration models. These concepts will now be specified.

25.2.1.3 Competing Risks Duration Models

Let us assume that the number of states K is strictly greater than 2 ($K>2$) and that, for each spell, there exists $(K-1)$ independent latent random variables, denoted $U^\star_{k,\ell}$ ($k \neq x_{\tau_{\ell-1}};\ k \in E$). Each random variable $U^*_{k,\ell}$ represents the latent sojourn duration in state $x_{\tau_{\ell-1}}$ before a transition to state k ($k \neq x_{\tau_{\ell-1}}$) during the $\ell-th$ spell of the process.

The observed sojourn duration u_ℓ is the minimum of these $(K-1)$ latent durations:

$$u_\ell \;=\; \inf_{k\neq x_{\tau_{\ell-1}}} \left\{u^*_{k,\ell}\right\} \tag{25.13}$$

Then, for any $\tau_{\ell-1} \in \omega$:

$$S(u \mid y_0,\ldots,y_{\ell-1};z;\theta) = \prod_{\substack{k=1\\k\neq j}}^{K} S(u,k \mid y_0,\ldots,y_{\ell-1};z;\theta) \tag{25.14}$$

where $S(u,k \mid y_0,\ldots,y_{\ell-1};z;\theta) = \Pr(U^*_{k,\ell} \geq u \mid y_0,\ldots,y_{\ell-1};z)$ is the conditional survival function of the sojourn duration in state $x_{\tau_{\ell-1}}$ before a transition to state k during the $\ell-th$ spell of the process, given the history of the process up to time $\tau_{\ell-1}$.

Let $g(u,k \mid y_0,\ldots,y_{\ell-1};z;\theta)$ be the conditional density function of the latent sojourn duration in state $x_{\tau_{\ell-1}}$ before a transition to state k, and $h_k(u \mid y_0,\ldots,y_{\ell-1};z;\theta)$ the associated conditional hazard function. Then we have the relations:

$$h_k(u \mid y_0,\ldots,y_{\ell-1};z;\theta) = \frac{g(u,k \mid y_0,\ldots,y_{\ell-1};z;\theta)}{S(u,k \mid y_0,\ldots,y_{\ell-1};z;\theta)} \tag{25.15}$$

and

$$S(u,k \mid y_0,\ldots,y_{\ell-1};z;\theta) = \exp\left(-\int_0^u h_k(t \mid y_0,\ldots,y_{\ell-1};z;\theta)\,\mathrm{d}t\right) \tag{25.16}$$

Let us remark that (25.14) and (25.16) imply:

$$S(u \mid y_0,\ldots,y_{\ell-1};z;\theta) = \exp\left(-\int_0^u \sum_{k\neq x_{\tau_{\ell-1}}} h_k(t \mid y_0,\ldots,y_{\ell-1};z;\theta)\,\mathrm{d}t\right) \quad (25.17)$$

Thus the conditional density function of the observed sojourn duration in state j during the $\ell-th$ spell of the process, given that this spell starts at time $\tau_{\ell-1}$ and ends at time $\tau_{\ell-1}+u$ by a transition to state k, is:

$$\begin{aligned} f(u,k \mid y_0,\ldots y_{\ell-1};z;\theta) &= h_k(u \mid y_0,\ldots,y_{\ell-1};z;\theta), \\ &\times \exp\left(-\int_0^u \sum_{\substack{k'=1 \\ k'\neq x_{\tau_{\ell-1}}}}^{K} h_{k'}(t \mid y_0,\ldots,y_{\ell-1};z;\theta)\,\mathrm{d}t\right) \end{aligned} \quad (25.18)$$

This is the likelihood contribution of the $\ell-th$ spell when this spell is not right-censored (i.e. when $\tau_\ell = \tau_{\ell-1}+u \leq \tau_e$). When the $\ell-th$ spell lasts more than $\tau_e - \tau_{\ell-1}$, the contribution of this spell to the likelihood function is:

$$S(\tau_e - \tau_{\ell-1} \mid y_0,\ldots,y_{\ell-1};z;\theta) = \Pr(U_\ell > \tau_e - \tau_{\ell-1} \mid y_0,\ldots,y_{\ell-1};z)$$

25.2.1.4 Multi-spells Multi-states Duration Models

These models are the extension of the preceding independent competing risks model, which treats the case of a single spell (the $\ell-th$ spell) with multiple destinations. In the multi-spells multi-states model, the typical likelihood contribution has the following form:

$$\mathscr{L}(\theta) = \prod_{\ell=1}^{n+1} f(y_\ell \mid y_0,\ldots,y_{\ell-1};z;\theta) \quad (25.19)$$

where $f(y_\ell \mid y_0,\ldots,y_{\ell-1};\theta)$ is the conditional density of Y_ℓ given $Y_0 = y_0, Y_1 = y_1,\ldots,Y_{\ell-1} = y_{\ell-1}, Z = z$ and θ is a vector of parameters. Definition (25.18) implies that:

$$\begin{aligned} \mathscr{L}(\theta) = \prod_{\ell=1}^{n} & f(\tau_\ell - \tau_{\ell-1}, x_{\tau_\ell} | y_0,\ldots,y_{\ell-1};z;\theta) \\ & \times S_{n+1}(\tau_e - \tau_n | y_0,\ldots,y_n;z;\theta) \end{aligned} \quad (25.20)$$

The last term of the right-hand side product in (25.20) is the contribution of the last observed spell, which is right-censored. References for a general presentation

of labor market transition econometric models can be found in surveys by Flinn and Heckman (1982a, b, 1983a) or in the textbook by Lancaster 1990a.

25.2.2 Non-parametric and Parametric Estimation

25.2.2.1 Non-parametric Estimation

The Kaplan-Meier Estimator

In the elementary duration model, a nonparametric estimator of the survivor function can be obtained using the Kaplan-Meier estimator for right-censored data. Let us suppose that we observe I sample paths (i.i.d. realizations of the process X_t) with the same past history $\omega[\tau_0, \tau_{n-1}]$. Let $I^\star$ be the number of sample paths such that $\tau_{n,i} \leq T_2$ and $I - I^\star$ the number of sample paths for which the n-th spell duration is right-censored, i.e. $\tau_{n,i} > T_2$, i denoting here the index of the process realization ($i = 1, \ldots, I$). If $\tau_{n,1}, \ldots, \tau_{n,I^\star}$ are the $I^\star$ ordered transition dates from state $X_{\tau_{n-1}}$ (i.e. $\tau_{n,1} \leq \ldots \leq \tau_{n,I^\star} \leq T_2$), the Kaplan-Meier estimator of the survivor function $S_n(t \mid \omega[\tau_0, \tau_{n-1}])$ is:

$$\hat{S}_n(t \mid \omega[\tau_0, \tau_{n-1}]) = \prod_{i:\tau_{n,i} \leq t} \left(1 - \frac{d_i}{r_i}\right)$$
$$i = 1, \ldots, I^\star,\ t \in]\tau_{n-1}, T_2] \tag{25.21}$$

where r_i is the number of sample paths for which the transition date from state $X_{\tau_{n-1}}$ is greater than or equal to $\tau_{n,i}$ and d_i is the number of transition times equal to $\tau_{n,i}$. An estimator for the variance of the survivor function estimate is given by the Greenwood's formula:

$$\begin{aligned} &\text{Var}\left[\hat{S}_n(t \mid \omega[\tau_0, \tau_{n-1}])\right] \\ &\quad \simeq \left\{\hat{S}_n(t \mid \omega[\tau_0, \tau_{n-1}])\right\}^2 \times \sum_{i:\tau_{n,i} \leq t} \frac{d_i}{r_i(r_i - d_i)} \end{aligned} \tag{25.22}$$

This estimator allows to implement nonparametric tests for the equality of the survivor functions of two different subpopulations (such as the Savage and log-rank tests).

In the case of multiple destinations (i.e. competing risks models), we must restrict the set of sample paths indexed by $i \in \{1, \ldots, I^\star\}$ to the process realizations experiencing transitions from the state $X_{\tau_{n-1}}$ to some state k ($k \neq X_{\tau_{n-1}}$). Transitions to another state than k are considered as right-censored durations. If we set $X_{\tau_{n-1}} = j$, then the Kaplan-Meier estimator of the survivor function $S_{jk}(t \mid \omega[\tau_0, \tau_{n-1}])$ is given by the appropriate application of formula (25.21), and an estimator of its variance is given by formula (25.22).

The Aalen Estimator

The function $H_\ell(u \mid \omega[\tau_0, \tau_{\ell-1}])$, defined in (25.12) and giving the integrated hazard function of the sojourn duration in the $\ell - th$ spell, can be estimated nonparametrically using the Aalen estimator (Aalen, 1978):

$$\hat{H}_\ell(u \mid \omega[\tau_0, \tau_{\ell-1}]) = \sum_{i:\tau_{\ell-1} \leq \tau_{\ell,i} < u} \frac{d_i}{r_i} \tag{25.23}$$

$\hat{H}_\ell(u \mid \omega[\tau_0, \tau_{\ell-1}])$ is an unbiased estimator of $H_\ell(u \mid \omega[\tau_0, \tau_{\ell-1}])$, and an estimator of its variance is given by:

$$var\left[\hat{H}_\ell(u \mid \omega[\tau_0, \tau_{\ell-1}])\right] = \sum_{i:\tau_{\ell-1} \leq \tau_{\ell,i} < u} \frac{d_i}{r_i(r_i - d_i)} \tag{25.24}$$

In the competing risks model, (25.12) is equivalent to:

$$\begin{aligned} -\log\, S_{jk}(u \mid \omega[\tau_0, \tau_{\ell-1}]) &= \textstyle\int_0^u h_{jk}(t \mid \omega[\tau_0, \tau_{\ell-1}])\, dt \\ &= H_{jk}(u \mid \omega[\tau_0, \tau_{\ell-1}]) \end{aligned} \tag{25.25}$$

where $H_{jk}(u \mid \omega[\tau_0, \tau_{\ell-1}])$ is the integrated intensity (or hazard) function for a transition from state j to state k ($k \neq j$) during the $\ell - th$ spell of the process, and given the past history $\omega[\tau_0, \tau_{\ell-1}]$ of the process. The Aalen estimator of this function can be derived from the formula (25.24) by considering indexes i corresponding to transitions from state j to state k during the $\ell - th$ spell of the process; indexes corresponding to other types of transition from state j are now considered as right-censored durations. The Aalen estimator can be used to implement nonparametric tests for the equality of two or more transition intensities corresponding to distinct transitions.

25.2.2.2 Specification of Conditional Hazard Functions

The Markov Model

In a markovian model, the hazard functions $h_k(t \mid y_0, \ldots, y_{\ell-1}; z; \theta)$ depend on t, on states $x_{\tau_{\ell-1}}$ and on k, but are independent of the previous history of the process. More precisely:

$$h_k(t \mid y_0, \ldots, y_{\ell-1}; z; \theta) = h_k(t \mid x_{\tau_{\ell-1}}; z; \theta), \quad k \neq x_{\tau_{\ell-1}} \tag{25.26}$$

and

$$h_j(t \mid y_0, \ldots, y_{\ell-1}; z; \theta) = 0, \text{ if } j = x_{\tau_{\ell-1}}$$

When the Markov model is time-independent, it is said to be time-homogeneous. In this case:

$$h_k(t \mid x_{\tau_{\ell-1}}; z; \theta) = h_k(x_{\tau_{\ell-1}}; z; \theta) = h_{x_{\tau_{\ell-1}},k}(z; \theta) \ , \ k \neq x_{\tau_{\ell-1}}, \ \forall t \in \mathbb{R}^+ \tag{25.27}$$

The particular case of a continuous-time markovian model observed in discrete-time will be extensively treated in the following subsection (this Chapter). Let us now consider two simple examples of markovian processes.

Example 25.1. Consider the case of a time-homogeneous markovian model with two states ($K = 2$) and assume that:

$$h_k(t \mid x_{\tau_{\ell-1}}; \theta) = \begin{cases} \alpha \ \text{ if } x_{\tau_{\ell-1}} = 1 \text{ and } k = 2 \\ \beta \ \text{ if } x_{\tau_{\ell-1}} = 2 \text{ and } k = 1 \\ 0 \text{ otherwise} \end{cases} \tag{25.28}$$

with $\theta = (\alpha, \beta)$. The parameter $\alpha > 0$ is the instantaneous rate of transition from state 1 (for instance, the employment state) to state 2 (for instance, the unemployment state). Reciprocally, $\beta > 0$ is the instantaneous rate of transition from state 2 to state 1.

Durations of employment (respectively, unemployment) are independently and identically distributed according to an exponential distribution with parameter α (respectively, with parameter β). If $p_1(t_0)$ and $p_2(t_0)$ denote occupation probabilities of states 1 and 2 at time t_0 respectively, then occupation probabilities at time t $(t > t_0)$ are respectively defined by:

$$\begin{aligned} p_1(t) &= \frac{\beta}{\alpha+\beta} + \left\{ p_1(t_0) - \frac{\beta}{\alpha+\beta} \right\} e^{-(\alpha+\beta)t} \\ p_2(t) &= \frac{\alpha}{\alpha+\beta} + \left\{ p_2(t_0) - \frac{\alpha}{\alpha+\beta} \right\} e^{-(\alpha+\beta)t} \end{aligned} \tag{25.29}$$

Let $(p_1^\star, \ p_2^\star)$ denote the stationary probability distribution of the process. Then it is easy to verify from (25.29) that:

$$p_1^\star = \frac{\beta}{\alpha+\beta} \ \text{ and } \ p_2^\star = \frac{\alpha}{\alpha+\beta} \tag{25.30}$$

In the economic literature, there are many examples of stationary job search models generating such a markovian time-homogeneous model with two states (employment and unemployment): see, for instance, the survey by Mortensen 1986. Extensions to three-states models (employment, unemployment and out-of-labor-force states) have been considered, for example, by Flinn and Heckman (1982a) and Burdett et al. (1984a, b). Markovian models of labor mobility have been estimated, for instance, by Tuma and Robins (1980), Flinn and Heckman (1983b), Mortensen and Neuman (1984), Olsen, Smith and Farkas, (1986) and Magnac and Robin (1994).

Example 25.2. Let us consider now the example of a non-homogeneous markovian model with two states (employment and unemployment, respectively denoted

1 and 2). Let us assume that the corresponding conditional hazard functions verify

$$h_k(t \mid x_{\tau_{\ell-1}};\theta) = \begin{cases} h_2(t;\theta) \text{ if } x_{\tau_{\ell-1}} = 1 \text{ and } k = 2 \\ h_1(t;\theta) \text{ if } x_{\tau_{\ell-1}} = 2 \text{ and } k = 1 \\ 0 \text{ otherwise} \end{cases} \tag{25.31}$$

Let $p(0) = (p_1(t_0), p_2(t_0))'$ denote the initial probability distribution at time t_0. The distribution of state occupation probabilities at time t, denoted $p(t) = (p_1(t), p_2(t))'$, is given by:

$$\begin{aligned} p_1(t) = \exp\left\{-\int_{t_0}^{t} [h_1(s;\theta) + h_2(s;\theta)]\, \mathrm{d}s\right\} \\ \times \left[p_1(t_0) + \int_{t_0}^{t} h_1(s;\theta)\, \exp\left\{\int_{t_0}^{s} (h_1(u;\theta) + h_2(u;\theta))\, \mathrm{d}u\right\}\, \mathrm{d}s\right] \end{aligned} \tag{25.32}$$

and $p_2(t) = 1 - p_1(t)$ (see Chesher and Lancaster, 1983).

Non-homogeneous markovian models are often used to deal with processes mainly influenced by the individual age at the transition date. For example, let us consider a transition process $\{X_t\}_{t\geq 0}$ with state-space $E = \{1,2,3\}$, and for which the time scale is the age (equal to A_t at time t). If the origin date of the process (i.e. the date of entry into the labor market) is denoted A_{τ_0} for a given individual, then a realization of the process $\{X_t\}_{t\geq 0}$ over the period $[A_{\tau_0},\ \tau_e]$ is depicted in Fig. 25.2.

Now let us suppose that transition intensities at time t depend only on the age attained at this time and are specified such as:

$$\begin{aligned} h_k\left(t \mid y_0, \ldots, y_{\ell-1}; A_{\tau_0}; \theta\right) &= h_k(A_t; x_{\tau_{\ell-1}}; \theta) \\ &= \exp\left(\alpha_{x_{\tau_{\ell-1}},k} + \beta_{x_{\tau_{\ell-1}},k} A_t\right) \end{aligned} \tag{25.33}$$

where $\alpha_{j,k}$ and $\beta_{j,k}$ ($j,k \in E \times E$ and $k \neq j$) are parameters to be estimated. In formula (25.33), the individual index is omitted for simplifying notations. By noting

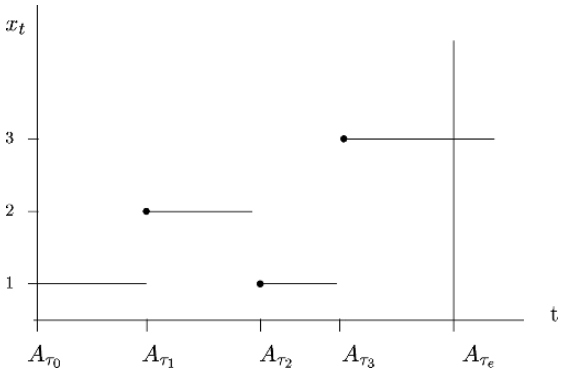

Fig. 25.2 A realization of a non-homogeneous markovian model

that:

$$A_t = A_{\tau_{\ell-1}} + (A_t - A_{\tau_{\ell-1}}) = A_{\tau_{\ell-1}} + u_{t\ell} \tag{25.34}$$

where $u_{t\ell}$ denotes the time already spent in the $\ell - th$ spell at date t, it is possible to write again transition intensities as:

$$h_k(t \mid y_0, \ldots, y_{\ell-1}; A_{\tau_0}; \theta) = \exp\left(\alpha_{x_{\tau_{\ell-1}},k} + \beta_{x_{\tau_{\ell-1}},k} A_{\tau_{\ell-1}} + \beta_{x_{\tau_{\ell-1}},k} u_{t\ell}\right) \tag{25.35}$$

and to deduce the survivor function of the sojourn duration in the $\ell - th$ spell which has the form:

$$S(u \mid y_0, \ldots, y_{\ell-1}; A_{\tau_0}; \theta)$$

$$= \exp\left\{ - \sum_{k \neq x_{\tau_{\ell-1}}} \int_{A_{\tau_{\ell-1}}}^{A_{\tau_{\ell-1}}+u} \exp(\alpha_{x_{\tau_{\ell-1}},k} + \beta_{x_{\tau_{\ell-1}},k} A_{\tau_{\ell-1}} + \beta_{x_{\tau_{\ell-1}},k} u_{t\ell}) \, \mathrm{d}t \right\} \tag{25.36}$$

where $\ell \geq 1$. By setting $u_{t\ell} = t - A_{\tau_{\ell-1}}$ in expression (25.36), it follows that:

$$S\{u \mid y_0, \ldots, y_{\ell-1}; A_{\tau_0}; \theta\} =$$

$$\exp\left(- \sum_{k \neq x_{\tau_{\ell-1}}} \frac{\exp(\alpha_{x_{\tau_{\ell-1}},k})}{\beta_{x_{\tau_{\ell-1}},k}} \left[\exp(\beta_{x_{\tau_{\ell-1}},k} (A_{\tau_{\ell-1}} + u)) - \exp(\beta_{x_{\tau_{\ell-1}},k} A_{\tau_{\ell-1}})\right] \right) \tag{25.37}$$

if $\beta_{x_{\tau_{\ell-1}},k} \neq 0$. Then the likelihood contribution of the $\ell - th$ spell beginning at age $A_{\tau_{\ell-1}}$ with a transition to state $x_{\tau_{\ell-1}}$ and ending at age A_{τ_ℓ} with a transition to state x_{τ_ℓ} is:

$$\begin{aligned} L_\ell &= f(A_{\tau_\ell} - A_{\tau_{\ell-1}}, x_{\tau_\ell} \mid y_0, \ldots, y_{\ell-1}; A_{\tau_0}; \theta) \\ &= h_{x_{\tau_\ell}}(\tau_\ell \mid y_0, \ldots, y_{\ell-1}; A_{\tau_0}; \theta) \, S(A_{\tau_\ell} - A_{\tau_{\ell-1}} \mid y_0, \ldots, y_{\ell-1}; A_{\tau_0}; \theta) \\ &= \exp\left(\alpha_{x_{\tau_{\ell-1}},k} + \beta_{x_{\tau_{\ell-1}},k} A_{\tau_\ell}\right) \\ &\quad \times \exp\left(- \sum_{k' \neq x_{\tau_{\ell-1}}} \frac{\exp(\alpha_{x_{\tau_{\ell-1}},k'})}{\beta_{x_{\tau_{\ell-1}},k'}} \left[\exp(\beta_{x_{\tau_{\ell-1}},k'} A_{\tau_\ell}) - \exp(\beta_{x_{\tau_{\ell-1}},k'} A_{\tau_{\ell-1}})\right] \right) \end{aligned} \tag{25.38}$$

Non-homogeneous markovian models of transitions between employment and unemployment have been estimated, for example, by Ridder 1986 and Trivedi and Alexander (1989).

Semi-Markov Models

In semi-Markov models, hazard functions depend only on the currently occupied state (denoted $x_{\tau_{\ell-1}}$ for spell ℓ), on the destination state (denoted k), on the sojourn duration in state $x_{\tau_{\ell-1}}$ and on the time of entry into the currently occupied state. If the spell corresponding to the currently occupied state is the $\ell-th$ spell of the process, then hazard functions of the semi-Markov model have two alternative representations:

$$h_k(t \mid y_0,\ldots,y_{\ell-1};\theta) = h_k(t \mid \tau_{\ell-1};x_{\tau_{\ell-1}};\theta) \tag{25.39}$$

or

$$h_k(u \mid y_0,\ldots,y_{\ell-1};\theta) = h_k(u \mid \tau_{\ell-1};x_{\tau_{\ell-1}};\theta) \tag{25.40}$$

where $u = t - \tau_{\ell-1}$ is the time already spent in the current state (i.e. in the $\ell-th$ spell of the process). When the hazard functions do not depend on the date $\tau_{\ell-1}$ of the last event, but depend only on the time already spent in the current state, then the semi-Markov model is said to be time-homogeneous. In this case, hazard functions defined in (25.40) are such that:

$$h_k(u \mid \tau_{\ell-1};x_{\tau_{\ell-1}};\theta) = h_k(u \mid x_{\tau_{\ell-1}};\theta),\ u \in \mathbb{R}^+ \tag{25.41}$$

In this model, the mean duration of a sojourn in state $x_{\tau_{\ell-1}}$ can be calculated using definitions of hazard and survivor functions, and thus it is given by:

$$E(U_\ell \mid x_{\tau_{\ell-1}};\theta) = \int_0^\infty u\, S(u \mid x_{\tau_{\ell-1}};\theta)\left\{\sum_{k\neq x_{\tau_{\ell-1}}} h_k(u \mid x_{\tau_{\ell-1}};\theta)\right\} \mathrm{d}u \tag{25.42}$$

where U_ℓ is the random variable representing the duration of a spell ℓ and

$$S(u \mid x_{\tau_{\ell-1}};\theta) = \exp(-\int_0^u \sum_{k\neq x_{\tau_{\ell-1}}} h_k(s \mid x_{\tau_{\ell-1}};\theta)\, \mathrm{d}s) \tag{25.43}$$

This conditional expectation can be obtained using the following property:

$$E(U_\ell \mid x_{\tau_{\ell-1}};\theta) = \int_0^\infty S(u \mid x_{\tau_{\ell-1}};\theta)\, \mathrm{d}u \tag{25.44}$$

(see, for instance, Klein and Moeschberger (2003)). Semi-markovian models of transition between two or three states have been estimated by Flinn and Heckman (1982b), Burdett, Kiefer and Sharma (1985), Bonnal, Fougère and Sérandon (1997), and Gilbert, Kamionka and Lacroix (2001).

25.2.3 Unobserved Heterogeneity

Here heterogeneity is supposed to cover individual observable and unobservable characteristics. Once again, we will omit the individual index.

25.2.3.1 Correlation Between Spells

Let us assume that the conditional model is time-homogeneous semi-markovian and

$$h_k(u \mid y_0,\ldots,y_{\ell-1};z;v;\theta) = h_k(u_\ell \mid x_{\tau_{\ell-1}};z;v_{x_{\tau_{\ell-1}},k};\theta_{x_{\tau_{\ell-1}},k}) \tag{25.45}$$

where v is a vector of individual unobserved heterogeneity terms and θ is the vector of parameters to be estimated.

Let $h_k(u_\ell \mid x_{\tau_{\ell-1}};z;v_{x_{\tau_{\ell-1}},k};\theta_{x_{\tau_{\ell-1}},k})$ denote the conditional hazard function for the sojourn duration in the $\ell-th$ spell of the participation process, when the currently occupied state is state $x_{\tau_{\ell-1}}$ and the destination state is k. Here z is a vector of exogenous variables, possibly time-dependent, $v_{(j,k)}$ is an heterogeneity random term, which is unobserved, and λ_{jk} is a vector of parameters. The preceding hazard function is often supposed to be equal to:

$$h_k(u_\ell \mid x_{\tau_{\ell-1}};z,v_{x_{\tau_{\ell-1}},k},\theta_{x_{\tau_{\ell-1}},k}) = \exp\left[\varphi(z;u_\ell;\theta_{x_{\tau_{\ell-1}},k}) + v_{x_{\tau_{\ell-1}},k}\right] \tag{25.46}$$

Several assumptions can be made concerning the unobserved random terms $v_{j,k}$. Firstly, $v_{j,k}$ can be supposed to be specific to the transition from j to k, so

$$v_{j,k} \neq v_{j',k'} \quad \text{for any } (j,k) \neq (j',k').$$

It can be also specific to the origin state, in which case:

$$v_{j,k} = v_j \quad \text{for any } k \neq j.$$

Finally, $v_{j,k}$ can be supposed to be independent of states j and k and thus to be fixed over time for each individual, i.e.

$$v_{j,k} = v \quad \text{for any } (j,k) \in E\times E,\ k \neq j.$$

This last assumption will be made through the remaining part of our presentation. Let us remark that a fixed heterogeneity term is sufficient to generate some correlation between spells durations. If we assume that v has a probability density function with respect to the Lebesgue measure denoted $g(v \mid \alpha)$, where α is a parameter, then we can deduce that the marginal survivor function of the sojourn duration in the $\ell-th$ spell of the process, when current state is $x_{\tau_{\ell-1}}$, has the form:

$$S(u_\ell \mid x_{\tau_{\ell-1}}; z; \theta_{x_{\tau_{\ell-1}}}) = \int_{D_G} S(u_\ell \mid x_{\tau_{\ell-1}}; z; v; \theta_{x_{\tau_{\ell-1}}})\; g(v \mid \alpha)\, \mathrm{d}v$$

$$= \int_{D_G} \exp\left\{ -\exp(v) \int_0^{u_\ell} \left(\sum_{k \neq x_{\tau_{\ell-1}}} \exp(\varphi(z; t; \theta_{x_{\tau_{\ell-1}}, k}) \right) \mathrm{d}t \right\} g(v \mid \alpha)\, \mathrm{d}v \quad (25.47)$$

where $\theta_{x_{\tau_{\ell-1}}} = \left\{ (\theta_{x_{\tau_{\ell-1}}, k})_{k \neq x_{\tau_{\ell-1}}}, \alpha \right\}$ and D_G is the support of the probability distribution of the random variable v.

Such formalizations of heterogeneity have been used for estimation purposes by Heckman and Borjas (1980), Butler et al. (1986), (1989), Mealli and Pudney (1996), Bonnal, Fougère and Sérandon (1997), Gilbert, Kamionka and Lacroix (2001), and Kamionka and Lacroix (2003).

• Example. To illustrate the treatment of unobserved heterogeneity in transition processes, let us consider a realization of a two state time-homogeneous Markov process. More precisely, let us assume that this realization generates a complete spell in state 1 over the interval $[0, \tau_1]$ and a right-censored spell in state 2 over the interval $[\tau_1, \tau_e[$. Transition intensities between the two states are given by:

$$h_k(t \mid x_{\tau_{\ell-1}}; v_{x_{\tau_{\ell-1}}}; \lambda_{x_{\tau_{\ell-1}}}) = \lambda_{x_{\tau_{\ell-1}}} + v_{x_{\tau_{\ell-1}}} \quad (25.48)$$

where $k \in \{1,2\}$, $\lambda_{x_{\tau_{\ell-1}}} > 0$ and $t \in \mathbb{R}^+$, λ_1 and λ_2 are two positive parameters, and v_1 and v_2 are two random variables supposed to be exponentially distributed with a density function $g(v \mid \alpha) = \alpha\, \exp(-\alpha\, v)$, $\alpha > 0$. We want to deduce the likelihood function for this realization of the process when v_1 and v_2 are supposed to be spell-specific and independent ($v_1 \neq v_2$ and $v_1 \perp\!\!\!\perp v_2$) or fixed over time ($v_1 = v_2 = v$). In the first case ($v_1 \neq v_2$ and $v_1 \perp\!\!\!\perp v_2$), the conditional likelihood function is:

$$L_v(\lambda) = f(\tau_1, x_{\tau_1} \mid x_0; v; \lambda)\; S(\tau_e - \tau_1 \mid x_{\tau_1}; v; \lambda),$$

$$= (\lambda_1 + v_1)\; \exp\{-(\lambda_1 + v_1)\tau_1\}\; \exp\{-(\lambda_2 + v_2)(\tau_e - \tau_1)\} \quad (25.49)$$

where $v = (v_1, v_2)'$, $\lambda = (\lambda_1, \lambda_2)'$, $x_0 = 1$ and $x_{\tau_1} = 2$. Because v_1 and v_2 are unobserved, we must deal with the following marginalized likelihood function:

$$L(\alpha; \lambda) = \int_0^\infty \int_0^\infty L(v_1, v_2, \lambda_1, \lambda_2)\; g(v_1 \mid \alpha)\; g(v_2 \mid \alpha)\; \mathrm{d}v_1\; \mathrm{d}v_2$$
$$= f(\tau_1, x_{\tau_1} \mid x_0; \alpha; \lambda)\; S(\tau_e - \tau_1 \mid x_1; \alpha; \lambda) \quad (25.50)$$

where

$$f(\tau_1, x_{\tau_1} \mid x_0; \alpha; \lambda) = \exp(-\lambda_1\; \tau_1) \left(\frac{\alpha}{\tau_1 + \alpha} \right) \left(\lambda_1 + \frac{1}{\tau_1 + \alpha} \right)$$

$$\text{and } S(\tau_e - \tau_1 \mid x_{\tau_1}; \alpha; \lambda) = \exp(-\lambda_2\; (\tau_e - \tau_1)) \left(\frac{\alpha}{(\tau_e - \tau_1) + \alpha} \right) \quad (25.51)$$

are the marginalized density and survivor functions of sojourn durations τ_1 and $(\tau_e - \tau_1)$ in the first and second spells respectively.

When the heterogeneity term is fixed over time ($v_1 = v_2 = v$), then the marginal likelihood contribution is:

$$\begin{aligned} L(\alpha,\lambda) &= \int_0^\infty (\lambda_1+v)\, \exp\{-(\lambda_1\tau_1+\lambda_2(\tau_e-\tau_1)+v\,\tau_e)\}\,\alpha\, \exp(-\alpha\, v)\, \mathrm{d}v, \\ &= \exp\{-\lambda_1\tau_1 - \lambda_2(\tau_e-\tau_1)\}\,\frac{\alpha}{\alpha+\tau_e}\left\{\lambda_1 + \frac{\alpha}{\alpha+\tau_e}\right\} \end{aligned} \tag{25.52}$$

which is obviously not equal to the product of the marginalized density and survivor functions of the sojourn durations in the first and second spells as in the case where $v_1 \neq v_2$. ⋆

Now, let us assume that there exists a function ψ defining a one-to-one relation between v and some random variable ν, such as:

$$v = \psi(\nu, \alpha) \tag{25.53}$$

For instance, ψ can be the inverse of the c.d.f. for v, and ν can be uniformly distributed on $[0,1]$. Then:

$$S(u_\ell \mid x_{\tau_{\ell-1}}; z; \theta_{x_{\tau_{\ell-1}}}) = \int_0^1 S(u_\ell \mid x_{\tau_{\ell-1}}; z; \psi(\nu,\alpha); \theta_{x_{\tau_{\ell-1}}})\, \phi(\nu)\, \mathrm{d}\nu \tag{25.54}$$

where $\phi(.)$ is the density function of ν. The marginal hazard function for the sojourn in the $\ell - th$ spell can be deduced from (25.54) as:

$$h(u_\ell \mid x_{\tau_{\ell-1}}; z; \theta_{x_{\tau_{\ell-1}}}) = -\frac{d}{\mathrm{d}u_\ell} S(u_\ell \mid x_{\tau_{\ell-1}}; z; \theta_{x_{\tau_{\ell-1}}}) \tag{25.55}$$

Using definitions (25.54) and (25.55), the individual contribution to the likelihood function can be easily deduced and maximized with respect to θ, either by usual procedures of likelihood maximization if the integrals (25.40) and (25.41) can be easily calculated, or by simulation methods (see, e.g., Gouriéroux and Monfort, 1997) in the opposite case.

For instance, let us consider the case of a semi-markovian model where the individual heterogeneity term is fixed over time, i.e. $v_{j,k} = v$ for any $(j,k) \in E \times E$. From (25.20) and (25.46)–(25.47), the typical likelihood contribution in the present case is:

$$\begin{aligned} L_v(\theta) = &\prod_{\ell=1}^{n} h_{x_{\tau_\ell}}\left(\tau_\ell - \tau_{\ell-1} \mid x_{\tau_{\ell-1}}; z; v; \theta_{x_{\tau_{\ell-1}}, x_{\tau_\ell}}\right) \\ &\times \prod_{\ell=1}^{n+1} \exp\left\{-\int_{\tau_{\ell-1}}^{\tau_\ell} \sum_{k \neq x_{\tau_{\ell-1}}} h_k(t \mid x_{\tau_{\ell-1}}; z; v; \theta_{x_{\tau_{\ell-1}},k})\, \mathrm{d}t\right\} \end{aligned} \tag{25.56}$$

with $\tau_{n+1} = \tau_e$ by convention. Using relation (25.53), the marginalized likelihood contribution obtained by integrating out v is:

$$\mathscr{L}(\theta) = \int_0^1 L_{\psi(v,\alpha)}(\theta)\ \phi(v)\ \mathrm{d}v \tag{25.57}$$

When the integral is not analytically tractable, simulated ML estimators of parameters α and $(\theta_{jk})_{k\neq j}$ can be obtained by maximizing the following simulated likelihood function with respect to α and $(\theta_{jk})_{k\neq j}$:

$$L_N(\theta) = \frac{1}{N}\sum_{n=1}^{N} L_{\psi(v_n,\alpha)}(\theta) \tag{25.58}$$

where v_n is drawn from the distribution with density function $\phi(.)$, which must be conveniently chosen (for asymptotic properties of these estimators, see Gouriéroux and Monfort, (1997)).

25.2.3.2 Correlation Between Destination States

Let us assume that the conditional hazard function for the transition into state k is given by the expression

$$h_k(u \mid y_0,\ldots,y_{\ell-1};z;v;\lambda) = h_k^0(u;\gamma)\ \varphi(y_0,\ldots,y_{\ell-1};z;\beta)\ \zeta_k \tag{25.59}$$

where $\varphi(.)$ is a positive function depending on the exogenous variables and the history of the process, ζ_k an unobserved heterogeneity component specific to the individual ($\zeta_k > 0$), β and γ are vectors of parameters, $h_k^0(u;\gamma)$ is a baseline hazard function for the transition to state k ($k \in \{1,\ldots,K\}$). Let us assume that (see Gilbert et al., 2001)

$$\zeta_k = \exp(a_k\ v_1 + b_k\ v_2) \tag{25.60}$$

where a_k and b_k are parameters such that $a_k = \mathbb{I}[k \geq 2]$ for $k = 1,\ldots,K$ and $b_1 = 1$. The latent components v_1 and v_2 are assumed to be independently and identically distributed with a p.d.f. denoted $g(v;\alpha)$, where α is a parameter and $v_s \in D_G$, $s = 1,2$.

In this two factor loading model, the correlation between $\log(\zeta_k)$ and $\log(\zeta_{k'})$, $\rho_{k,k'}$, is given by the expression

$$\rho_{k,k'} = \frac{a_k\ a_{k'} + b_k\ b_{k'}}{\sqrt{a_k^2 + b_k^2}\sqrt{a_{k'}^2 + b_{k'}^2}} \tag{25.61}$$

where $k,k' = 1,\ldots,K$. The contribution to the conditional likelihood function of a given realization of the process $w = (y_1,\ldots,y_n,y_{n+1})$ is:

$$\mathscr{L}(\theta)=\int_{D_G}\int_{D_G}\prod_{\ell=1}^{n+1} f(y_\ell|y_0,\ldots,y_{\ell-1};z;v_1,v_2;\lambda)\; g(v_1;\alpha)\; g(v_2;\alpha)\,\mathrm{d}v_1\,\mathrm{d}v_2 \tag{25.62}$$

where

$$f(u,k\mid y_0,\ldots,y_{\ell-1};z;v_1,v_2;\lambda)=h_k(u\mid y_0,\ldots,y_{\ell-1};z;v_1,v_2;\lambda)^{\delta_k} \times\exp\left\{-\int_0^u\sum_{j\neq x_{\tau_{\ell-1}}} h_j(t\mid y_0,\ldots,y_{\ell-1};z;v_1,v_2;\lambda)\,\mathrm{d}t\right\} \tag{25.63}$$

and the conditional hazard function is given by expression (25.59). The exponent δ_k is equal to 1 if $k\in\{1,\ldots,K\}$, and to 0 otherwise. λ is a vector of parameters and $\theta=(\alpha,\lambda)$. As the last spell is right-censored, the corresponding contribution of this spell is given by the survivor function

$$f(y_{n+1}|y_0,\ldots,y_n;z;v_1,v_2;\lambda)=\exp\left\{-\int_0^{u_{n+1}}\sum_{j\neq x_{\tau_n}} h_j(t|y_0,\ldots,y_n;z;v_1,v_2;\lambda)\,\mathrm{d}t\right\} \tag{25.64}$$

where $y_{n+1}=(u_{n+1},0)$ (state 0 corresponds to right-censoring).

Bonnal et al. (1997) contains an example of a two factor loading model. Lindeboom and van den Berg (1994), Ham and Lalonde (1996) and Eberwein et al. (1997), (2002) use a one factor loading model in order to correlate the conditional hazard functions. A four factor loading model has been proposed by Mealli and Pudney (2003). Let us remark that, in the case of bivariate duration models, association measures were studied by Van den Berg 1997. Discrete distributions of the unobserved heterogeneity component can be alternatively used (see, for instance, Heckman and Singer (1984) Gritz 1993, Baker and Melino (2000)).

This way to correlate the transition rates using a factor loading model is particularly useful for program evaluation on nonexperimental data. In this case, it is possible to characterize the impact on the conditional hazard functions of previous participation to a program by taking into account selectivity phenomena at entry into the program.

25.3 Markov Processes Using Discrete-Time Observations

The econometric literature on labor mobility processes observed with discrete-time panel data makes often use of two elementary stochastic processes describing individual transitions between a finite number of participation states.

The first one is the continuous-time Markov chain, whose parameters can be estimated through the quasi-Newton (or scoring) algorithm proposed by Kalbfleisch and Lawless (1985). This kind of model allows to calculate stationary probabilities

of state occupation, the mean duration of sojourn in a given state, and the intensities of transition from one state to another.

A main difficulty can appear in this approach: in some cases the discrete-time Markov chain cannot be represented by a continuous-time process. This problem is known as the embeddability problem which has been surveyed by Singer and Spilerman (1976a), b) and Singer (1981, 1982). However, some non-embeddable transition probability matrices can become embeddable after an infinitesimal modification complying with the stochastic property. This suggests that the embeddability problem can be due to sampling errors.

Geweke et al., (1986a) established a bayesian method to estimate the posterior mean of the parameters associated with the Markov process and some functions of these parameters, using a diffuse prior defined on the set of stochastic matrices. Their procedure allows to determine the embeddability probability of the discrete-time Markov chain and to derive confidence intervals for its parameters under the posterior.

The second frequently used modelization incorporates a very simple form of heterogeneity among the individuals: this is the mover-stayer model, which was studied in the discrete-time framework by Frydman 1984, Sampson 1990 and Fougère and Kamionka (2003). The mover-stayer model is a stochastic process mixing two Markov chains. This modelling implies that the reference population consists of two types of individuals: the "stayers" permanently sojourning in a given state, and the "movers" moving between states according to a non-degenerate Markov process.

These two modelizations will be successively studied in the following subsection.

25.3.1 The Time-Homogeneous Markovian Model

Let us consider a markovian process $\{X_t,\ t \in \mathbb{R}^+\}$ defined on a discrete state-space $E = \{1,\ldots,K\}$, $K \in \mathbb{N}$, with a transition probability matrix $P(s,t)$ with entries $p_{j,k}(s,t)$, $(j,k) \in E \times E$, $0 \leq s \leq t$, where:

$$p_{j,k}(s,t) = \Pr\{X_t = k \mid X_s = j\} \tag{25.65}$$

and $\sum_{k=1}^{K} p_{j,k}(s,t) = 1$. If this markovian process is time-homogeneous, then:

$$p_{j,k}(s,t) = p_{j,k}(0,t-s) \equiv p_{j,k}(t-s),\ \ 0 \leq s \leq t \tag{25.66}$$

or equivalently:

$$P(s,t) = P(0,t-s) \equiv P(t-s),\ \ 0 \leq s \leq t \tag{25.67}$$

This implies that transition intensities defined by:

$$h_{j,k} = \lim_{\Delta t \downarrow 0} p_{j,k}(t,t+\Delta t)/\Delta t,\ \ \Delta t \geq 0,\ (j,k) \in E \times E,\ j \neq k \tag{25.68}$$

are constant through time, i.e.:

$$h_k(t \mid x_{\tau_{\ell-1}}; \theta) = h_{j,k}(t \mid \theta) = h_{j,k}\ ,\ t \geq 0,\ (j,k) \in E \times E,\ j \neq k \tag{25.69}$$

where $x_{\tau_{\ell-1}} = j$. These transition intensities are equal to the hazard functions previously defined in (25.26) and (25.27). The $K \times K$ transition intensity matrix, which is associated to the time-homogeneous markovian process $\{X_t, t \in \mathbb{R}^+\}$, is denoted Q and has entries:

$$q(j,k) \ = \ \begin{cases} h_{j,k} \in \mathbb{R}^+ \text{ if } j \neq k,\ (j,k) \in E \times E \\ -\sum\limits_{\substack{m=1 \\ m \neq j}}^{K} h_{j,m} \leq 0 \ if\ j = k,\ j \in E \end{cases} \tag{25.70}$$

Let us denote $\mathbb{Q}$ the set of transition intensity matrices, i.e. the set of $(K \times K)$ matrices with entries verifying the conditions (25.70). It is well known (cf. Doob (1953, p. 240 and 241) that the transition probability matrix over an interval of length T can be written:

$$P(0,T) \ = \ \exp(QT),\ T \in \mathbb{R}^+ \tag{25.71}$$

where $\exp(A) = \sum_{k=0}^{\infty} A^k/k!$ for any $K \times K$ matrix A.

Main properties of the time-homogeneous markovian process $\{X_t,\ t \in \mathbb{R}^+\}$ with state-space E, are the following:

- sojourn times in state j ($j \in$ E) are positive random variables, which are exponentially distributed with parameter $-q(j,j)$:

$$u_j \sim \exp(-q(j,j)),\ j = 1, \ldots, K \tag{25.72}$$

  with $E[u_j] = \text{var}[u_j]^{1/2} = -q(j,j)^{-1}$,
- the probability of a transition to state k given that the process is currently in state j ($k \neq j$) is independent of the sojourn time in state j, and is found to be:

$$r_{j,k} = -q(j,k)/q(j,j),\ k \neq j,\ (j,k) \in E \times E \tag{25.73}$$

- if the time-homogeneous Markov process $\{X_t\}$ is ergodic, its equilibrium (or limiting) probability distribution is denoted $P^* = (p_1^*, \ldots, p_K^*)'$ and defined as the unique solution to the linear system of equations:

$$Q'P^* = 0\ \text{, with}\ \sum_{i=1}^{K} p_i^* = 1 \tag{25.74}$$

25.3.1.1 Maximum Likelihood Estimator of the Matrix P Using Discrete-Time (Multiwave) Panel Data

Let us suppose now that we observe η independent realizations of the process $\{X_t\}$ at equally spaced times $T_0,\ T_1,\ldots,\ T_L\ (L>1)$ such as: $T_\ell - T_{\ell-1} = T,\ \ell = 1,\ldots,L$. Let us denote:

- $n_{j,k}(\ell)$ the number of individuals who were in state j at time $T_{\ell-1}$ and who are in state k at time T_ℓ,
- $n_j(\ell-1)$ the number of individuals who were in state j at time $T_{\ell-1}$. Maximizing the conditional likelihood function given the initial distribution at T_0:

$$\begin{aligned} L(P(0,T)) &= \prod_{\ell=1}^{L} \prod_{j,k=1}^{K} \left\{ p_{j,k}(T_{\ell-1}, T_\ell) \right\}^{n_{j,k}(\ell)} \\ &= \prod_{j,k=1}^{K} \left\{ p_{j,k}(0,T) \right\}^{\Sigma_{\ell=1}^{L} n_{j,k}(\ell)} \end{aligned} \tag{25.75}$$

with $\sum_{k=1}^{K} p_{j,k}(0,T) = 1$, gives the (j,k) entry of the MLE $\widehat{P}(0,T)$ for $P(0,T)$:

$$\widehat{p}_{j,k}(0,T) = \left(\sum_{\ell=1}^{L} n_{j,k}(\ell) \right) / \left(\sum_{\ell=1}^{L} n_j(\ell-1) \right) \tag{25.76}$$

(see Anderson and Goodman, 1957). If the solution $\widehat{Q}$ to the equation:

$$\widehat{P}(0,T) = \exp(\widehat{Q}T),\ T>0 \tag{25.77}$$

belongs to the set $\mathbb{Q}$ of intensity matrices, then $\widehat{Q}$ is a MLE estimator for Q. Nevertheless, two difficulties may appear:[3]

- the (25.77) can have multiple solutions $\widehat{Q} \in \mathbb{Q}$: this problem is known as the aliasing problem;[4]
- none of the solutions $\widehat{Q}$ to the equation (25.77) belongs to the set $\mathbb{Q}$ of intensity matrices; in that case, the probability matrix $\widehat{P}(0,T)$ is said to be non-embeddable with a continuous-time Markov process.

25.3.1.2 Necessary Conditions for Embeddability

The unique necessary and sufficient condition for embeddability was given by Kendall, who proved that, when $K=2$, the transition matrix $\widehat{P}(0,T)$ is embeddable

[3] A detailed analysis of these problems is developed in papers by Singer and Spilerman (1976 a and b).

[4] The aliasing problem has also been considered by Phillips (1973).

if and only if the trace of $\widehat{P}(0,T)$ is strictly greater than 1. When $K \geq 3$, only necessary conditions are known; they are the following:[5]
1st necessary condition (Chung, 1967):

- if $\widehat{p}_{j,k}(0,T) = 0, then\ \widehat{p}_{j,k}^{(n)}(0,T) = 0,\ \forall n \in \mathbb{N}$, where $\widehat{p}_{j,k}^{(n)}(0,T)$ is the entry (j,k) of the matrix $[\widehat{P}(0,T)]^n$,
- if $\widehat{p}_{j,k}(0,T) \neq 0$, then $\widehat{p}_{j,k}^{(n)}(0,T) \neq 0,\ \forall n \in \mathbb{N}$;

2nd necessary condition (Kingman, 1962): $\det\left[\widehat{P}(0,T)\right] > 0$,
3rd necessary condition (Elfving, 1937):

- no eigenvalue λ_i of $\widehat{P}(0,T)$ can satisfy $\mid \lambda_i \mid = 1$, other than $\lambda_i = 1$;
- in addition, any negative eigenvalue must have even algebraic multiplicity;

4th necessary condition (Runnenberg, 1962): the argument of any eigenvalue λ_i of $\widehat{P}(0,T)$ must satisfy:

$$\left(\frac{1}{2} + \frac{1}{K}\right)\Pi \leq \arg(\log\ \lambda_i) \leq \left(\frac{3}{2} - \frac{1}{K}\right)\Pi$$

This last condition plays an important role in the remainder of the analysis.

25.3.1.3 Resolving the Equation $\hat{P}(0,T) = \exp(\hat{Q}T)$

The proof of the following theorem can be found in Singer and Spilerman (1976a):

If $\widehat{P}(0,T)$ has K distinct[6] *eigenvalues $(\lambda_1,\ldots,\lambda_K)$ and can be written $\widehat{P}(0,T) = A \times D \times A^{-1}$, where $D = diag(\lambda_1,\ldots,\lambda_K)$ and the eigenvector corresponding to λ_i $(i = 1,\ldots,K)$ is contained in the i_th column of the $(K \times K)$ matrix A, then:*

$$\log(\widehat{P}(0,T)) = \widehat{Q}T = A \times \begin{pmatrix} \log_{k_1}(\lambda_1) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \log_{k_K}(\lambda_K) \end{pmatrix} \times A^{-1} \qquad (25.78)$$

where $\log_{k_i}(\lambda_i) = \log \mid \lambda_i \mid + (\arg\lambda_i + 2k_i\Pi)i, k_i \in \mathbb{Z}$, is a branch of the logarithm of λ_i, when $\lambda_i \in \mathscr{C}$.[7]

---

[5] Singer and Spilerman (1976a) and Geweke, Marshall and Zarkin (1986b) survey this problem.

[6] The case of repeated eigenvalues arises very rarely in empirical applications. For its treatment, the reader can consult Singer and Spilerman (1976a), p. 19–25).

[7] Let us recall that the logarithmic function is multiple valued in the complex set $\mathscr{C}$. If $z = a + ib$ $(z \in \mathscr{C})$, then: $\log_k(z) = \log \mid z \mid + i(\theta + 2k\Pi)$, $k \in \mathbb{Z}$, with $\mid z \mid = \sqrt{a^2 + b^2}$, and $\theta = \arg(z) = \tan^{-1}(b/a)$. Each value for k generates a distinct value for $\log(z)$, which is called a branch of the logarithm.

Since (25.77) has as many solutions $\widehat{Q}$ as there are combinations of the form $(\log_{k_1}(\lambda_1),\ldots,\log_{k_K}(\lambda_K))$, the number of these solutions is infinite when the matrix $\widehat{P}(0,T)$ has at least two complex conjugate eigenvalues. However, an important implication of the fourth necessary condition for embeddability is that only finitely many branches of $\log(\widehat{P}(0,T))$ need to be checked for membership in $\mathbb{Q}$. Indeed, this condition implies:

$$\forall\lambda_i,\ -L_i(K)\leq k_i\leq U_i(K) \tag{25.79}$$

$$\text{where}\quad U_i(K)=\text{intpt}\left|\frac{\log\mid\lambda_i\mid\tan\{(\frac{1}{2}+\frac{1}{K})\Pi\}-\mid\arg\lambda_i\mid}{2\Pi}\right|$$

$$L_i(K)=\text{intpt}\left|\frac{\log\ \mid\lambda_i\mid\tan\{(\frac{3}{2}-\frac{1}{K})\Pi\}-\mid\arg\lambda_i\mid}{2\Pi}\right|$$

the function "intpt" being the integer part of a real number. So the number of branches of λ_i which must be computed is equal to $L_i(K)+U_i(K)+1$, the last one corresponding to the main branch (with $k_i=0$). Then the number of solutions $\widehat{Q}$ that must be examined for membership in $\mathbb{Q}$ is denoted $k^*(\widehat{P})$ and is equal to:

$$k^*(\widehat{P})=\begin{cases}\displaystyle\prod_{j=1}^{v}\{L_j(K)+U_j(K)+1\} & \text{if } v\geq 1\\ 1 & \text{if } v=0\end{cases} \tag{25.80}$$

where v denotes the number of complex conjugate eigenvalue pairs of the matrix $\widehat{P}(0,T)$. Let us remark that:

- for a real eigenvalue, only the principal branch of the logarithm must be examined: other branches (with $k_i\neq 0$) correspond to complex intensity matrices $\widehat{Q}$;
- each element of a complex conjugate eigenvalue pair has the same number of candidate branches (see (25.79)); moreover, only combinations of branches involving the same k_i in each element of the pair must be computed; all others correspond to complex intensity matrices; this fact explains why the calculation of $k^*(\widehat{P})$ is based on the number of complex conjugate eigenvalue pairs, and why the number of branches that need to be checked for each pair j is equal to $L_j(K)+U_j(K)+1$ rather than $\{L_j(K)+U_j(K)+1\}^2$.

If (25.77) has only one solution $\widehat{Q}\in\mathbb{Q}$, this solution is the MLE for the intensity matrix of the homogeneous continuous-time Markov process $\{X_t,\ t\in\mathbb{R}^+\}$; an estimator for the asymptotic covariance matrix of $\widehat{Q}$ has been given by Kalbfleisch and Lawless (1985).

25.3.1.4 The Scoring Procedure

Kalbfleisch and Lawless (1985) have proposed to maximize with respect to θ the conditional likelihood function (25.75), i.e.

$$L(\theta) = \prod_{i,j=1}^{K} \left\{ \exp(QT) \right\}_{(i,j)}^{\sum_{\ell=1}^{L} n_{i,j}(\ell)}, \quad Q \in \mathbb{Q} \tag{25.81}$$

through a scoring algorithm. In this expression, $\{\exp(QT)\}_{i,j}$ is the entry (i,j) of the matrix $\exp(QT) = P(0,T)$ and θ is the vector of extra diagonal elements of the matrix Q ($\theta \equiv \theta(Q)$). If it is assumed that matrix Q has K distinct eigenvalues, denoted $(d_1, \cdots, d_K)$, matrices Q and $P(0,T)$ can be written as:

$$\begin{aligned} Q &= A\, D_Q A^{-1} = A \text{ diag } (d_1, \cdots, d_K) A^{-1} \\ \text{and } P(0,T) &= \exp(QT) = A \exp(D_Q T) A^{-1} \\ &= A \text{ diag}(e^{d_1 T}, \cdots, e^{d_K T}) A^{-1} = A \text{ diag}(\lambda_1, \cdots, \lambda_K) A^{-1} \end{aligned} \tag{25.82}$$

These formulae lead to a convenient expression of the score (or gradient) vector, which is:

$$S(\theta) = \left\{ \frac{\partial \log L(Q)}{\partial q_{k\ell}} \right\} = \left\{ \sum_{i,j=1}^{K} \sum_{\ell=1}^{L} n_{i,j}(\ell) \frac{\partial \{\exp(QT)\}_{(i,j)} / \partial q_{k\ell}}{\{\exp(QT)\}_{(i,j)}} \right\} \tag{25.83}$$

where

$$\begin{aligned} \frac{\partial \{\exp(QT)\}}{\partial q_{k\ell}} = \sum_{s=1}^{\infty} \left(\frac{\partial Q^s}{\partial q_{k\ell}} \right) \frac{T^s}{s!} &= \sum_{s=1}^{\infty} \sum_{r=0}^{s-1} Q^r \frac{\partial Q}{\partial q_{k\ell}} \cdot Q^{s-1-r} \cdot \frac{T^s}{s!} \\ &= A V_{k\ell} A^{-1} \end{aligned}$$

the matrix

$$V_{k\ell} = \sum_{s=1}^{\infty} \sum_{r=0}^{s-1} D_Q^r \left(A^{-1} \frac{\partial Q}{\partial q_{k\ell}} A \right) D_Q^{s-1-r} \frac{T^s}{s!} \text{ having elements:}$$

$$\begin{cases} (G_{k\ell})_{(i,j)} \dfrac{e^{d_i t} - e^{d_j t}}{d_i - d_j}, i \neq j, \\ (G_{k\ell})_{(i,j)} t\, e^{d_i t} \quad , i = j, \end{cases}$$

where $(G_{k\ell})_{(i,j)}$ is the entry (i,j) of the matrix $G_{k\ell} = A^{-1} \dfrac{\partial Q}{\partial q_{k\ell}} A$.

The information matrix, which has the form

$$E\left[-\frac{\partial^2 \log L(\theta)}{\partial q_{k\ell}\partial q_{k'\ell'}}\right] = \left\{\sum_{\ell=1}^{L}\sum_{i,j=1}^{K}\frac{E[N_i(\ell-1)]}{p_{i,j}(0,T)}\,\frac{\partial p_{i,j}(0,T)}{\partial q_{k\ell}}\,\frac{\partial p_{i,j}(0,T)}{\partial q_{k'\ell'}}\right\} \tag{25.84}$$

(see Kalbfleisch and Lawless ((1985), p. 864), is estimated by:

$$M(\theta) = \left\{\sum_{\ell=1}^{L}\sum_{i,j=1}^{K}\frac{n_i(\ell-1)}{p_{i,j}(0,T)}\,\frac{\partial p_{i,j}(0,T)}{\partial q_{k\ell}}\,\frac{\partial p_{i,j}(0,T)}{\partial q_{k'\ell'}}\right\} \tag{25.85}$$

The iterative formula for the scoring algorithm being:

$$\theta_{n+1} = \theta_n + M(\theta_n)^{-1}S(\theta_n)$$

where $n \geq 0$ and an initial value $\theta_0 = \theta(Q_0)$ is still to be chosen. Two cases must be considered (the case with multiple solutions in Q is excluded):

- equation (25.77) admits only one solution for $\widehat{Q}$ and this solution belongs to the set $\mathbb{Q}$ of transition intensity matrices: $\hat{Q}$ is the MLE of the transition matrix Q of the time-homogeneous markovian process, and the matrix $M(\theta(\widehat{Q}))^{-1}$ gives a consistent estimate of the covariance matrix of $\hat{\theta} = \theta(\widehat{Q})$;
- the unique solution $Q_0 = \widehat{Q}$ to (25.77) doesn't belong to the set $\mathbb{Q}$; however, it may exist matrices $\tilde{P}(0,T) = \exp(\tilde{Q}T)$ "close" to $\widehat{P}(0,T)$ and which are embeddable, i.e. such that $\tilde{Q} \in \mathbb{Q}$; in this case, the scoring algorithm of Kalbfleisch and Lawless (1985) can be applied to the maximization of the likelihood (25.81) subject to the constraint $Q \in \mathbb{Q}$; this constraint can be directly introduced into the iterative procedure by setting

$$q_{i,j} = \begin{cases} \exp(a_{i,j}),\ a_{i,j} \in \mathbb{R}, j \neq i, (i,j) \in E \times E \\ q_{ii} = -\sum_{\substack{k=1 \\ k\neq i}}^{K} q_{ik}, i = j,\ i \in E \end{cases} \tag{25.86}$$

and the initial value Q_0 can be chosen to verify:

$$Q_0 = \underset{Q\in\,\mathbb{Q}}{\operatorname{argmin}} \parallel Q_0 - \widehat{Q} \parallel \tag{25.87}$$

where $\widehat{Q} = \frac{1}{T}\log\widehat{P}(0,T)$.

25.3.1.5 Bayesian Inference

Geweke, Marshall and Zarkin (1986a) have developed a bayesian approach for statistical inference on Q (and functions of Q) by using a diffuse prior on the set of stochastic matrices. This approach can be justified by two arguments:

- when the MLE of Q is on the parameter set boundary, standard asymptotic theory cannot be applied any more; bayesian inference overcomes this difficulty: the posterior confidence interval for Q can be viewed as its asymptotic approximation;
- moreover, bayesian inference allows incorporating into the choice of the prior distribution some information external to the sample (for example, the distribution of sojourn durations in each state).

Let us denote P_K the set of $(K \times K)$ stochastic matrices, i.e. $P_k = \left\{P \in M_{K,K} : \forall i,j \in E, p_{i,j} \geq 0 \text{ and } \sum_{j=1}^{K} p_{i,j} = 1\right\}$, P_K^* the set of $(K \times K)$ embeddable stochastic matrices, i.e. $P_K^* = \{P \in M_{K,K} : P \in P_K \text{ and } \exists Q \in \mathbb{Q}, P(O,T) = \exp(QT), T > 0\}$. For any $P \in P_K^*, k^*(P)$ denotes the number of combinations of the form (25.78) belonging to $\mathbb{Q}$ and verifying (25.77). Now let us consider a prior distribution on $P \in P_K$, denoted $\mu(P)$, a prior distribution on Q, denoted $h_k(P)$ and verifying $\sum_{k=1}^{k^*(P)} h_k(P) = 1$ for $P \in P_K$, and a $\mathbb{R}$-valued function of interest denoted $g(Q)$. If the posterior embeddability probability of P is defined as:

$$\Pr(P \in P_K^* \mid N) = \frac{\int_{P_K^*} L(P;N)\mu(P)\mathrm{d}P}{\int_{P_K} L(P;N)\mu(P)\mathrm{d}P} > 0 \tag{25.88}$$

then the expectation of $g(Q)$ is equal to

$$E[g(Q) \mid N, P \in P_K^*] = \frac{\int_{P_K^*} \sum_{k=1}^{k^*(P)} h_k(P) g[Q_k(P)] L(P;N)\mu(P)\mathrm{d}P}{\int_{P_K^*} L(P;N)\mu(P)\mathrm{d}P} \tag{25.89}$$

where the entry (i,j) of the matrix N is $\sum_{\ell=1}^{L} n_{i,j}(\ell), L(P;N)$ is the likelihood function and $Q_k(P)$ is the transition intensity matrix corresponding to the k-th combination of logarithms of the eigenvalues of matrix P. The function of interest $g(Q)$ can be, for example, $g(Q) = q_{i,j}, (i,j) \in E \times E$, or:

$$g(Q) = E\left\{(q_{i,j} - E(q_{i,j} \mid N; P \in P_K^*))^2 \mid N; P \in P_K^*\right\}$$

which is equivalent to:

$$g(Q) = E\left\{q_{i,j}^2 \mid N; P \in P_K^*\right\} - E^2\left\{q_{i,j} \mid N; P \in P_K^*\right\}$$

The embeddability probability for P and the first moment of $g(Q)$ may be computed using Monte-Carlo integration. This involves the choice of an importance function from which a sequence of matrices $\{P_i\} \in P_K$ can be easily generated (see Geweke et al., (1986a), for such a function). Now let us consider a function $J(P_i)$ such that $J(P_i) = 1$ if $P_i \in P_K^*$ and $J(P_i) = 0$ otherwise. If $\mu(P_i)$ is bounded above, then:

$$\lim_{I\to\infty} \frac{\sum_{i=1}^{I} J(P_i)L(P_i;N)\mu(P_i)/I(P_i)}{\sum_{i=1}^{I} L(P_i;N)\mu(P_i)/I(P_i)} \tag{25.90}$$

$$= \Pr[P \in P_K^* \mid N] \qquad \text{a.s.}$$

Moreover, if $H_k(P)$ is a multinomial random variable such that $\Pr[H_k(P) = 1] = h_k(P)$, and if $g(Q)$ is bounded above, then

$$\lim_{I\to\infty} \frac{\sum_{i=1}^{I} \sum_{k=1}^{k^*(P_i)} H_k(P_i)g[Q_k(P_i)]J(P_i)L(P_i;N)\mu(P_i)/I(P_i)}{\sum_{i=1}^{I} J(P_i)L(P_i;N)\mu(P_i)/I(P_i)}$$

$$= E[g(Q) \mid N; P \in P_K^*] \qquad \text{a.s.} \tag{25.91}$$

(see Geweke et al., (1986a), p. 658).

25.3.1.6 Tenure Records

Up to now we concentrated on the statistical analysis of discrete-time observations of an underlying continuous-time Markov process. The available information is sometimes richer than the one brought by discrete-time data, but not as complete as the one contained in continuous-time data. Indeed it can consist, for a given individual, in the joint sequence $\{(x_{T_\ell}, d_{T_\ell})\}_{\ell=0,\cdots,L}$ of occupied states $\{x_{T_\ell}\}_{\ell=0,\cdots,L}$ and of times $\{d_{T_\ell}\}_{\ell=0,\cdots,L}$ already spent in these states at distant observation times $\{T_\ell\}_{\ell=0,\cdots L}$. Such data have been studied in the continuous-time markovian framework by Magnac and Robin (1994), who proposed to call this kind of observations "tenure records". Figure 25.3 gives an example of a tenure record.

In this example, T_0, T_1, T_2 and T_3 are the exogenous survey dates. The process $\{X_t\}_{t\geq 0}$ is first observed to be in state $x_{T_0} = 1$ at time T_0: it occupies this state from date $(T_0 - d_0)$ on. It is then observed to be in state 3 at successive times T_1 and T_2. This state was entered at time $(T_1 - d_1) = (T_2 - d_2)$. Finally, the process is at time T_3 in state $x_{T_3} = 1$ from date $(T_3 - d_3)$ on. Indeed it is possible that a spell covers two survey dates, as it is the case for the second observed spell in the preceding example: obviously, the information collected in T_1 is redundant.

Let us remark that in tenure records data sets, any sojourn duration is right-censored with probability one. Typically, a tenure record consists of a sequence $\{x_{T_\ell}, d_\ell, t_\ell\}_{\ell=0,\cdots,L}$ with the convention $t_L = \infty$. The process $\{X_t\}_{t\geq 0}$ enters state x_{T_ℓ} at time $(T_\ell - d_\ell)$ and is observed to stay in this state for a duration greater than d_ℓ. Then the process is not observed (i.e. is truncated) during a period of length

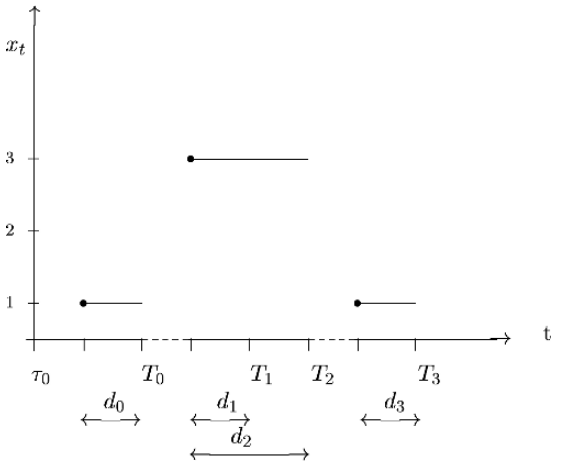

Fig. 25.3 An example of a tenure record

$t_\ell = (T_{\ell+1} - d_{\ell+1}) - T_\ell$. Let $h_{ij}(s,t)$ be the probability that the process $\{X_t\}$ enters state j at time t given that it was in state i at time $s (s < t)$. If $\{X_t\}$ is time-homogeneous markovian, then $h_{ij}(0, t-s) \equiv h_{ij}(t-s), s < t$. In this case, $h_{ij}(t)$ is equal to:

$$h_{ij}(t) = \sum_{\substack{k=1 \\ k \neq j}}^{K} p_{ik}(t)\, q_{kj},\ (i,j) \in E \times E \tag{25.92}$$

Consequently, the likelihood function for a tenure record $\{x_{T_\ell}, d_\ell, t_\ell\}_{\ell=0,\cdots,L}$ is the following:

$$\begin{aligned} L &= \left\{ \prod_{\ell=0}^{L-1} S(d_\ell \mid x_{T_\ell})\, h_{x_{T_\ell}, x_{T_{\ell+1}}}(t_\ell) \right\} S(d_L \mid x_{T_L}) \\ &= \exp(-\lambda_{x_{T_L}} d_L) \prod_{\ell=0}^{L-1} \left\{ \exp(-\lambda_{x_{T_\ell}} d_\ell) \sum_{\substack{k=1 \\ k \neq x_{T_{\ell+1}}}}^{K} \{\exp(Qt_\ell)\}_{(x_{T_\ell}, k)} \cdot q_{k, x_{T_{\ell+1}}} \right\} \end{aligned} \tag{25.93}$$

where $S(u \mid x_{T_\ell})$ is the survivor function of the sojourn duration in state x_{T_ℓ} and Q is the transition intensity matrix with entries:

$$Q(i,j) = \begin{cases} -\lambda_i = -\sum_{\substack{k=1 \\ k \neq i}}^{K} q_{ik}, \text{ if } j = i \\ \\ q_{ij}, \text{ if } j \neq i \end{cases}$$

Magnac and Robin (1994) show that tenure records allow to identify the intensity of transition from one state to the same state (for example, employment) when

within-state mobility is allowed (i.e. when a worker can directly move from one job to another). Discrete-time observations do not present this advantage.

For a treatment of incomplete records, particularly in presence of unobserved heterogeneity see, for instance, Kamionka 1998. Magnac et al. (1995) propose to use indirect inference to estimate the parameters of a transition model under a semi-Markov assumption in the context of a censoring mechanism.

25.3.2 The Mover-Stayer Model

25.3.2.1 MLE for the Discrete-Time Mover-Stayer Model

The mover-stayer model has been introduced by Blumen et al. 1955 for studying the mobility of workers in the labor market. Subsequently, Goodman (1961), Spilerman (1972) and Singer and Spilerman (1976a) have developed the statistical analysis of this model, essentially on the discrete-time axis. The mover-stayer model in discrete time is a stochastic process $\{X_\ell,\ \ell \in \mathbb{N}\}$, defined on a discrete state-space $E = \{1,\ldots,K\}$, $K \in \mathbb{N}$, and resulting from the mixture of two independent Markov chains; the first of these two chains, denoted $\{X^1_\ell, \ell \in \mathbb{N}\}$ is degenerate, i.e. its transition probability matrix is the identity matrix, denoted I. The other chain, denoted $\{X^2_\ell\ ,\ \ell \in \mathbb{N}\}$ is characterized by a non-degenerate transition matrix $M(s,u) = \| m_{i,j}(s,u) \|$, $i,j = 1,\ldots,K$, $0 \le s \le u$, where:

$$m_{i,j}(s,u) = \Pr\{X^2_u = j \mid X^2_s = i\},\ i,j \in E,\ s,u \in \mathbb{N},\ s \le u \tag{25.94}$$

and $\displaystyle\sum_{j=1}^{K} m_{i,j}(s,u) = 1.$

Moreover, the Markov chain $\{X^2_\ell, \ell \in \mathbb{N}\}$ is assumed to be time homogeneous, i.e.:

$$m_{i,j}(s,u) = m_{i,j}(0,u-s) \equiv m_{i,j}(u-s),\ \ 0 \le s \le u \tag{25.95}$$

which is equivalent to:

$$M(s,u) = M(0,u-s) \equiv M(u-s),\ \ 0 \le s \le u \tag{25.96}$$

Now let us assume that the mixed process $\{X_\ell, \ell \in \mathbb{N}\}$ is observed at fixed and equally distant times: $0,T,2T,\ldots,LT$, with $T > 0$ and $L \in \mathbb{N}$ $(L \ge 1)$. Transition probabilities for this process are given by the formulas:

$$p_{i,j}(0,kT) = \Pr[X_{kT} = j \mid X_0 = i],\ i,j \in E,\ k = 1,\ldots,L \tag{25.97}$$

$$= \begin{cases} (1-s_i)[m_{i,j}(T)]^{(k)} & \text{if } \ j \neq i \\ s_i + (1-s_i)[m_{i,i}(T)]^{(k)} & \text{if } \ j = i \end{cases}$$

where $[m_{i,j}(T)]^{(k)}$ is the entry (i,j) of the matrix $[M(T)]^k$, and $(s_i, 1-s_i)$, with $s_i \in [0,1]$, is a mixing measure for state $i \in E$. So, in the mover-stayer model, the

reference population is composed of two kinds of individuals: the "stayers", permanently sojourning in the same state, and the "movers", who move from one state to another according to the time-homogeneous Markov chain with transition probability matrix $M(s,u), s \leq u$. The proportion of "stayers" in state i $(i \in E)$ is equal to s_i.

The estimation of the transition matrix $M(0,T)$ and of the mixing measure s from a sample of N independent realizations of the process $\{X_\ell, \ell \in \mathbb{N}\}$, has been extensively treated by Frydman (1984) and then carried out by Sampson (1990). The method developed by Frydman relies on a simple recursive procedure, which will be rapidly surveyed. Formally, the form of the sample is:

$$\{X_{0(n)}, X_{T(n)}, X_{2T(n)}, \ldots, X_{LT(n)};\ 1 \leq n \leq N\}$$

where $X_{kT(n)}$ $(k = 0, \ldots, L)$ is the state of the process for the $n_$th realization at time kT, and $(L+1)$ is the number of equally spaced dates of observation.

Let us denote $n_{i_0,\ldots,i_{LT}}$ the number of individuals for which the observed discrete path is $(i_0, \ldots, i_{LT})$, $n_i(kT)$ the number of individuals in state i at time kT, $n_{ij}(kT)$ the number of individuals who are in state i at time $(k-1)T$ and in state j at time (kT), n_i the number of individuals who have a constant path,[8] i.e. $i_0 = i_T = \ldots = i_{LT} = i, i \in E$, $n_{ij} = \sum_{k=1}^{L} n_{ij}(kT)$ the total number of observed transitions from state i to state j, $n_i^* = \sum_{k=0}^{L-1} n_i(kT)$ the total number of visits to state i before time (LT), $\eta_i \geq 0$ the proportion of individuals initially (i.e. at date 0) in state i, $i \in E$, with $\sum_{i=1}^{K} \eta_i = 1$.

The likelihood function for the sample is (Frydman, 1984, p. 633):

$$L = \prod_{i=1}^{K} \eta_i^{n_i(0)} \prod_{i=1}^{K} L_i \tag{25.98}$$

where:

$$L_i = \{s_i + (1-s_i)[m_{ii}(0,T)]^L\}^{n_i} (1-s_i)^{n_i(0)-n_i} [m_{ii}(0,T)]^{n_{ii}-Ln_i} \times \prod_{\substack{k=1 \\ k \neq i}}^{K} [m_{ik}(0,T)]^{n_{ik}}$$

In this last expression, $n_i(0)$ is the number of individuals in state i at time 0, n_i is the number of individuals permanently observed in state i, $(n_i(0) - n_i)$ is the number of individuals initially in state i who experience at least one transition in the L following periods, n_{ik} is the total number of transitions from state i to state k. Maximizing the function (25.98) with respect to M and s subject to the constraints $s_i \geq 0$, $i \in E$, is equivalent to maximize the K expressions:

$$\mathscr{L}_i = \text{Log}\, L_i + \lambda_i s_i,\ i = 1, \ldots, K \tag{25.99}$$

[8] Among the individuals permanently sojourning in state i, we must distinguish the "stayers" from the "movers"; indeed, the probability that a "mover" is observed to be in state i at each observation point is strictly positive and equal to $\{m_{ii}(0,T)\}^L$.

for which the first-order derivatives relatively to s_i are:

$$\frac{\partial \mathscr{L}_i}{\partial s_i} = \frac{n_i\{1-[m_{ii}(0,T)]^L\}}{s_i+(1-s_i)[m_{ii}(0,T)]^L} - \frac{n_i(0)-n_i}{1-s_i} + \lambda_i = 0 \tag{25.100}$$

Two situations should be considered:
First case: If $s_i > 0$, then $\lambda_i = 0$ and:

$$s_i = \frac{n_i - n_i(0)[m_{ii}(0,T)]^L}{n_i(0)\{1-[m_{ii}(0,T)]^L\}} \tag{25.101}$$

As shown by Frydman (1984, p. 634–635), the ML estimators of transition probabilities m_{ij} (with fixed i, and j varying from 1 to K) are given by the recursive equation:

$$\widehat{m}_{ij}(0,T) = n_{ij}\{1-\widehat{m}_{ii}(0,T) - \sum_{\substack{k=1\\k\neq i}}^{j-1} \widehat{m}_{ik}(0,T)\} / \sum_{\substack{k=j\\k\neq i}}^{K} n_{ik},\ j\neq i,\ i,j\in E \tag{25.102}$$

To solve (25.102), it is necessary to begin by setting $j = 1$ if $i \neq 1$ and $j = 2$ if $i = 1$. Furthermore, $\widehat{m}_{ii}(0,T)$ is the solution, belonging to the interval $[0,1]$, to the equation:

$$\begin{aligned}&[n_i^* - Ln_i(0)][m_{ii}(0,T)]^{L+1} + [Ln_i(0)-n_{ii}][m_{ii}(0,T)]^L\\ &\quad +[Ln_i - n_i^*]m_{ii}(0,T) + (n_{ii}-Ln_i) \ = \ 0\end{aligned} \tag{25.103}$$

Frydman (1984) doesn't notice that $s_i \leq 0$ whenever $(\frac{n_i}{n_i(0)}) \leq [m_{ii}(0,T)]^L$, where $(n_i/n_i(0))$ is the proportion of individuals permanently observed in state i. In that case, the initial assumption $s_i > 0$ is violated, and it is necessary to consider the case where $s_i = 0$.
Second case: If $s_i = 0$, then:

$$\widehat{m}_{ij}(0,T) = n_{ij}/n_i^*,\ \forall\, i,j = 1,\ldots,K \tag{25.104}$$

This is the usual ML estimator for the probability of transition from i to j for a first-order Markov chain in discrete time (for example, see Anderson and Goodman (1957), or Billingsley (1961). A remark, which is not contained in the paper by Frydman (1984), must be made. It may appear that $Ln_i = n_{ii}$ (with $n_{ii} \neq 0$), which means that no transition from state i to any other distinct state is observed. This case arises when the number n_i of individuals permanently observed in state i is equal to the number $n_i(0)$ of individuals initially present in state i (if $n_i(0) \neq 0$). Then the estimation problem has two solutions:

- s_i=1 and m_{ii} is non-identifiable (see (25.101) and (25.103)),
- $s_i = 0$ and $m_{ii} = 1$.

The first solution corresponds to a pure model of "stayers" in state i, the second to a time-homogeneous Markov chain in which state i is absorbing. The mover-stayer

model, as a mixture of two Markov chains, is not appropriate any more for state i. When this case appears in the applied work, we propose to choose the solution $s_i = 0$ and $m_{ii} = 1$, especially for computing the estimated marginal probabilities of the form $\Pr[X_{kT} = i]$, $k = 0, \ldots, L,\ i = 1, \ldots, K$. The analytical expression of the estimated asymptotic covariance matrix for ML estimators $\widehat{M}$ and $\widehat{s}$ can be calculated using second derivatives of expression (25.99).

25.3.2.2 Bayesian Inference for the Continuous-Time Mover-Stayer Model

The mover-stayer model in continuous-time is a mixture of two independent Markov chains; the first one denoted $\{X_t^1,\ t \in \mathbb{R}^+\}$ has a degenerate transition matrix equal to the identity matrix I; the second one denoted $\{X_t^2,\ t \in \mathbb{R}^+\}$ has a non-degenerate transition matrix $M(s,t)$, $0 \leq s \leq t$, verifying over any interval of length T:

$$M(0,T) = \exp(QT), \quad T \in \mathbb{R}^+ \tag{25.105}$$

Setting $M(0,kT) = \|m_{i,j}(0,kT)\|$, we get:

$$P(0,kT) = \operatorname{diag}(s) + \operatorname{diag}(\mathbb{I}_K - s)\{\exp(QT)\}^K,\ T \geq 0,\ k = 1, \ldots, L \tag{25.106}$$

where $s = (s_1, \ldots, s_K)'$, $(\mathbb{I}_K - s) = (1 - s_1, \ldots, 1 - s_K)'$, and $\operatorname{diag}(x)$ is a diagonal matrix with vector x on the main diagonal. From the discrete-time ML estimators of stayers' proportions s and of the transition probability matrix $M(0,T)$, it is then possible to obtain the ML estimator of the intensity matrix Q by resolving (25.105) (see subsection 2.1 above). But, due to the possible problem of non-embeddability of the matrix $M(0,T)$, it could be better to adopt a bayesian approach, as the one proposed by Fougère and Kamionka (2003). This approach is summarized below.

Definition 25.1 To write the likelihood-function and the expected value under the posterior of some function of parameters, additional notation is needed. Let M_K be the space of $K \times K$ stochastic matrices:

$$M_K = \{M = \| m_{ij} \| \ :\ m_{ij} \geq 0,\ \forall i, j \in E \text{ and } \sum_{j=1}^{K} m_{ij} = 1,\ \forall\, i \in E\}.$$

Clearly, the transition probability matrix $M(0,T)$ belongs to M_K. Let $\mu(M,s)$ be a prior mapping $M_K \times [0,1]$ into $\mathbb{R}$ (the uniform prior will be used in the application). $\mu(M,s)$ is defined for $M \in M_K$ and for a vector of mixing measures $s = \{s_i\ ,\ i \in E\} \in [0,1]^K$. $[0,1]^K$ denotes the cartesian product of K copies of $[0,1]$. Let us denote $\mathbb{Q}$ the space of intensity matrices:

$$\mathbb{Q} = \{Q = \| q_{ij} \| \ :\ q_{ij} \geq 0,\ i, j \in E,\ i \neq j \text{ and } q_{ii} \leq 0, \forall i \in E\}.$$

If $M(0,T)$ is embeddable, there exists at least one matrix $Q \in \mathbb{Q}$ defined by the equation $M(0,T) = \exp(QT)$, where T is the number of time units between observations. Let M_K^* the space of embeddable stochastic matrices:

$$M_K^* = \{ M(0,T) \in M_K : \exists Q \in \mathbb{Q}, \exp(QT) = M(0,T)\}.$$

If $D_K = M_K \times [0,1]^K$ represents the parameters space for the model, then the space $D_K^* = M_K^* \times [0,1]^K$ denotes the set of embeddable parameters and $D_K^* \subset D_K$. As it was shown in subsection 2.1, the solution to $M(0,T) = \exp(QT)$ may not be unique: this is the **aliasing problem**.

Let us consider now the set of matrices $Q^{(k)} \in \mathbb{Q}$, solutions of the equation $Q^{(k)} = \log(M(0,T))/T$, for $k = 1, \ldots, B(M)$. $B(M)$ is the number of continuous-time underlying processes corresponding to the discrete-time Markov chain represented by $M(0,T) \in M_K$. We have $B(M) \in \mathbb{N}$ and $B(M) = 0$ if $M \notin M_K^*$. Denote $Q^{(k)}(M)$ the intensity matrix that corresponds to the $k-$th solution of log(M), $k = 1, \ldots, B(M)$. $Q^{(k)}(M)$, $1 \leq k \leq B(M)$, is a function defined for $M \in M_K^*$, $Q^{(k)}(M) \in Q$. Let $h^{(k)}(M)$ be a probability density function induced by a prior probability distribution on the $k-$th solution of the equation $M(0,T) = \exp(QT)$ when $M \in M_K^*$. By definition, $h^{(k)}(M)$ verifies $\sum_{k=1}^{B(M)} h^{(k)}(M) = 1$.

Let $g(Q,s)$ be a function defined for $(Q,s) \in \mathbb{Q} \times [0,1]^K$. This function is such that the evaluation of its moments (in particular, the posterior mean and the posterior standard deviation) is a question of interest. Thus, the posterior probability that the transition probability matrix M is embeddable has the form:

$$\Pr[(M,s) \in D_K^* \mid (N,n)] = \frac{\displaystyle\int_{D_K^*} L(M,s;N,n)\mu(M,s)\, d(M,s)}{\displaystyle\int_{D_K} L(M,s;N,n)\mu(M,s)\, d(M,s)} \tag{25.107}$$

Likelihood and Importance Functions

The likelihood function $L \equiv L(M,s;N,n)$ up to the initial distribution of the process $\{X(t),\ t \geq 0\}$ is

$$L \propto \prod_{i=1}^{K} L_i \tag{25.108}$$

where:

$$\begin{aligned} L_i = [\, s_i + (1-s_i) \times \{\exp(QT)\}_{ii}^{L}\,]^{n_i} \times (1-s_i)^{n_i(0)-n_i} \\ \times \{\exp(QT)\}_{ii}^{n_{ii}-Ln_i} \prod_{k\neq i,k=1}^{K} \{\exp(QT)\}_{ik}^{n_{ik}}, \end{aligned} \tag{25.109}$$

$\{\exp(QT)\}_{i,k}$ denoting the entry (i,k) of the $K \times K$ matrix $\exp(QT)$. If $\Pr[M \in M_K^* \mid N,n] > 0$, then

$$E[g(Q,s) \mid (N,n);(M,s) \in D_K^*] \tag{25.110}$$

$$= \frac{\displaystyle\int_{D_K^*} \sum_{k=1}^{B(M)} h^{(k)}(M)\; g(Q^{(k)}(M),s)\; L(M,s;N,n)\; \mu(M,s)\; d(M,s)}{\displaystyle\int_{D_K} L(M,s;N,n)\; \mu(M,s)\; d(M,s)}$$

In order to evaluate the integrals inside expressions (25.107) and (25.110), an adaptation of the Monte-Carlo method may be used because an analytical expression for $Q^{(k)}(M)$ or $B(M)$ when $K \geq 3$ has not been found yet. Let $I(M,s)$ be a probability density function defined for $(M,s) \in D_K$. $I(M,s)$ is the importance function from which a sequence $\{M_i,s_i\}$ of parameters will be drawn. We suppose that $I(M,s) > 0$ and that $\mu(M,s)$ and $g(Q,s)$ are bounded above.

Let $J(M)$ a function defined for $M \in M_K$:

$$J(M) = \begin{cases} 1 \text{ if } M \in M_K^* \\ 0 \text{ otherwise} \end{cases}$$

Then

$$\lim_{I \to +\infty} \frac{\displaystyle\sum_{i=1}^{I} J(M_i)\; L(M_i,s_i;N,n)\; \mu(M_i,s_i)/I(M_i,s_i)}{\displaystyle\sum_{i=1}^{I} L(M_i,s_i;N,n)\; \mu(M_i,s_i)/I(M_i,s_i)} \tag{25.111}$$

$$\stackrel{a.s}{=} \Pr[(M,s) \in D_K^* \mid N,n]$$

and

$$E[g(Q,s) \mid N,n;(M,s) \in D_K^*] \stackrel{a.s}{=}$$

$$\lim_{I \to +\infty} \frac{\displaystyle\sum_{i=1}^{I} \sum_{k=1}^{B(M)} \frac{h^{(k)}(M_i)\; g[Q^{(k)}(M_i),s_i] J(M_i)\; L(M_i,s_i;N,n)\; \mu(M_i,s_i)}{I(M_i,s_i)}}{\displaystyle\sum_{i=1}^{I} J(M_i)\; L(M_i,s_i;N,n)\; \mu(M_i,s_i)/I(M_i,s_i)} \tag{25.112}$$

where $\Pr[(M,s) \in D_K^* \mid N,n]$ is the probability under the posterior that the discrete-time Mover–Stayer model is embeddable with the continuous time one, and $E[g(Q,s) \mid N,n;(M,s) \in D_K^*]$ defines the posterior moments of the parameters' function of interest.

For a better convergence of estimators (25.111) and (25.112), $I(M,s)$ should be concentrated on the part of D_K where $L(M,s;N,n)$ is nonnegligible. For that purpose, if $\mu(M,s)$ is not concentrated on some part of the set D_K (that's the case when μ is uniform), $I(M,s)$ can be taken proportional to the likelihood $L(M,s;N,n)$. Because drawing (M,s) from $L(M,s;N,n)$ is difficult, Fougère and Kamionka (2003)

choose a normal expansion for $L(M,s;N,n)$ with mean the ML estimator $(\widehat{M},\widehat{s})$ and with covariance matrix the inverse of the information matrix estimated at $(\widehat{M},\widehat{s})$.

When $g(Q,s)$ and $\mu(M,s)$ are bounded above, the convergence of the estimator (25.112) is obtained almost surely. When the function $g(Q,s)$ does not verify this property (for instance, if we are interested in the estimation of q_{ij}), the convergence of the expression (25.112) relies on the existence of the posterior mean: $E[g(Q,s) \mid (M,s) \in D_K^*;N,n]$.

The covariance matrix V associated to $L(M,s;N,n)$ is block diagonal with blocks consisting of matrices V_i, $i=1,\ldots,K$, defined as:

$$V_i(M,s) = -E\left[\frac{\partial^2 Log(L_i(M,s;N,n))}{\partial\theta_k\partial\theta_l}\right]^{-1} = R_i(M,s)^{-1} \tag{25.113}$$

with $\theta_k,\theta_l = \begin{cases} m_{i,j}\ ,\ i,j\in E \\ s_i\ ,\ i\in E \end{cases}$ where $R_i(M,s)$ is the $i-$th diagonal block of the information matrix $R(M,s)$ associated to $L(M,s;N,n)$. Then a sequence of draws $\{(M_k,s_k)\}_{k=1,\ldots,I}$ can be generated according to the density of a multivariate normal distribution with mean (M,s) and covariance matrix $V(M,s)=R(M,s)^{-1}$. If we suppose that $V_{i-}=P_iP_i'$ is the Choleski's decomposition of the matrix V_{i-} obtained by dropping the last row and column of matrix V_i , and if $y^k\sim N(0_K,I_K)$, then

$$z^k = P_i y^k + \begin{pmatrix} s_i \\ m_{i1} \\ \vdots \\ m_{iK-1} \end{pmatrix} \sim N(\begin{pmatrix} s_i \\ m_{i1} \\ \vdots \\ m_{iK-1} \end{pmatrix}, V_{i-}) \tag{25.114}$$

Finally, we can obtain m_{iK} by setting $m_{iK}=1-\sum_{j=1}^{K-1} m_{i,j}$. Inside the procedure, s_i, $(m_{i,1},\ldots,m_{i,K})$, and V_i are estimated by their MLE, respectively $\hat{s}_i$, $(\hat{m}_{i,1},\ldots,\hat{m}_{i,K})$, and $\hat{V}_i$. For more details, see Fougère and Kamionka (2003).

Limiting Probability Distribution and Mobility Indices

The mobility of movers can be appreciated by examination of the mobility indices for continuous-time Markov processes proposed by Geweke et al. (1986b). For the movers process with intensity matrix Q, four indices of mobility can be considered:

$$\begin{aligned} M_1(Q) &= -\log[\det(M(0,T))]/K = -tr(Q)/K \\ M_2(Q) &= \sum_{i=1}^{K}\Pi_i^{(m)}\sum_{j=1}^{K} q_{ij}\mid i-j\mid \\ M_3(Q) &= -\sum_{j=1}^{K}\Pi_j^{(m)} q_{ij} \\ M_4(Q) &= -\Re e[\log(\lambda_2)] \end{aligned} \tag{25.115}$$

where:

- $\Pi_i^{(m)}$ is the equilibrium probability in state i for the movers, given by equation $Q'\pi_i^{(m)} = 0$, with $\Sigma_{i=1}^K \pi_i^{(m)} = 1$,
- the eigenvalues of the matrix $M(0,T)$ denoted by $\lambda_1, \ldots, \lambda_K$, are ordered so that $\mid \lambda_1 \mid \geq \ldots \geq \mid \lambda_K \mid$,
- $\Re e$ denotes the real part of the logarithm of the eigenvalue λ_2.

We can also define the equilibrium (or limiting) probability distribution for the mixed "mover-stayer" process $\{X_t, t \in \mathbb{R}^+\}$. For state i, the limiting probability, denoted π_i, is given by:

$$\pi_i = s_i \eta_i + \pi_i^{(m)} \sum_{j=1}^{K} (1 - s_j)\eta_j, \ i \in E \tag{25.116}$$

where:

- $\eta = \{\eta_i, i \in E\}$ is the initial probability distribution (i.e. at the date 0) for the process $\{X_t, t \in \mathbb{R}^+\}$,
- and $\pi_i^{(m)}$ is the limiting probability of "movers" in state i.

It is easily verified that, for a purely markovian process (one for which $s_i = 0, \ \forall i \in E$), the formula (25.116) becomes $\pi_i = \pi_i^{(m)}$. The mobility indices (25.115) and the limiting distribution (25.116) can be estimated using formula (25.112) and taking respectively $g(Q,s) = M_k(Q)$ $(1 \leq k \leq 4)$, or $g(Q,s) = \pi$.

Bayesian Inference Using Gibbs Sampling

The likelihood function of the sample X can be written

$$L(X|s,M,X_0) = \prod_{n=1}^{N} \sum_{k=1}^{2} \mathscr{L}(X_{(n)}|s,M,X_{o(n)},z_n{=}k) \ \Pr[z_n{=}k|s,M,X_{o(n)}]$$

where $\mathscr{L}$ is the conditional contribution of the individual n given the initial state $X_{o(n)}$ and the unobserved heterogeneity type z_n. z_n is an unobserved indicator taking the value 1 if the individual is a stayer or the value 2 if the individual is a mover.

The prior density on the parameter $\theta = (s,M)$ is assumed to be the product of the conjugate densities $\mu_1(s)$ and $\mu_2(M)$, where

$$\mu_1(M) = \prod_{j=1}^{K} \frac{\Gamma(a_j + b_j)}{\Gamma(a_j)\Gamma(b_j)} s_j^{a_j - 1}(1 - s_j)^{b_j - 1}$$

is the Dirichlet distribution with parameters $a_j > 0$, $b_j > 0$, $j = 1, \ldots, K$, and

$$\mu_2(M) = \prod_{i=1}^{K} \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_{ik}\right)}{\prod_{k=1}^{K} \Gamma(\alpha_{ik})} \prod_{i,j=1}^{K} m_{ij}^{\alpha_{ij}-1}$$

is the matrix beta distribution with parameter $\alpha_{ij} > 0$, $i, j = 1, \ldots, K$.

The conditional distribution of the unobserved type z_n is thus

$$z_n \mid \theta, X_{(n)} \sim B(1; p(X_{(n)}; \theta)) \tag{25.117}$$

where

$$p(X_n; \theta) = \frac{\mathscr{L}(X_{(n)} \mid s, M, X_{o(n)}, z_n = 1) \Pr[z_n = 1 \mid s, M, X_{o(n)}]}{\sum_{i=1}^{2} \mathscr{L}(X_{(n)} \mid s, M, X_{o(n)}, z_n = i) \Pr[z_n = i \mid s, M, X_{o(n)}]}$$

Combining the prior and the sample informations we obtain that

$$s_j \mid X, Z \sim \text{ Dirichlet}\left(a_j + \sum_{n=1}^{N} i_j^{(n)}(2 - z_n), b_j + \sum_{n=1}^{N} i_j^{(n)}(z_n - 1)\right) \tag{25.118}$$

$$M \mid X, Z \sim \text{ Matrix beta}\left(\alpha_{ik} + \sum_{n=1}^{N} (z_n - 1) N_{ik}^{(n)}; i, k = 1, \ldots, K\right) \tag{25.119}$$

The Gibbs sampling algorithm runs like this:
Initialization: Fix an initial value $\theta^{(0)} = (s^{(0)}, M^{(0)})$.
Update from $\theta^{(m)}$ to $\theta^{(m+1)}$ by doing :
1 - Generate $Z^{(m)}$ according to the conditional distribution (25.117), given $\theta = \theta^{(m)}$ and X;
2 - Generate $\theta^{(m+1)} = (s^{(m+1)}, M^{(m+1)})$ using the conditional distribution (25.118) and (25.119), given $Z = Z^{(m)}$ and X.

Under general regularity conditions and for m large enough, the resulting random variable $\theta^{(m)}$ is distributed according to the stationary posterior distribution $\mu(\theta \mid X)$. Draws from the stationary posterior distribution $\mu(\theta \mid X)$ may be used to obtain posterior estimates of θ using an expression similar to the one given by (25.112) (see Fougère and Kamionka, (2003)). Step one of the algorithm corresponds to a data augmentation step (see, Robert and Casella, (2002)).

25.4 Concluding Remarks

This chapter has introduced reduced-form models and statistical methods for analyzing longitudinal panel data on individual labor market transitions. The first section gave a very general presentation of methods concerning continuous-time

observations, while the second section focused on the treatment of discrete-time observations for continuous-time discrete-state processes.

Obviously, our survey did not intend to cover exhaustively a continuously and rapidly growing literature. Among subjects treated in this field of research, two topics seem to be especially important. The first one is the treatment of endogenous selection bias in dynamic populations (see Lancaster and Imbens, (1990), (1995), Lancaster, (1990b), Ham and Lalonde, (1996), and Fougère, Kamionka and Prieto, (2005)). Indeed, some sampling schemes for continuous-time discrete state space processes are such that the probability of being in the sample depends on the endogenous variable, i.e. being in a given state (for example, unemployment) at some date. Consequently inference from these endogenous samples requires specific statistical methods which have begun to be elaborated (see the papers quoted above). Another research area is the evaluation of the effect of public interventions such as employment and training programs. Here the main problem is knowing if these programs have a joint positive effect on earnings and employment rates of beneficiaries (see, for example, papers by Card and Sullivan, (1988), Heckman, (1990), Eberwein, Ham and Lalonde, (1997), Bonnal, Fougère and Sérandon, (1997), Heckman, Lalonde and Smith, (1999)). In order to avoid misleading results, this evaluation must take into account the selection biases induced simultaneously by the process of eligibility to the program and by the sampling scheme. Thus these two fields of research are very closely connected.

References

Aalen O.O., 1978, "Non parametric inference for a family of counting processes", The Annals of Statistics, Vol. 6, 701–726.

Amemiya T., 2001, "Endogenous Sampling in Duration Models", Monetary and Economic Studies, Vol. 19, No. 3, 77–96.

Anderson T.W. and Goodman L.A., 1957, "Statistical inference about Markov chains", Annals of Mathematical Statistics, Vol. 28, 89–110.

Baker M. and Melino A., 2000, "Duration dependence and nonparametric heterogeneity: a Monte Carlo study", Journal of Econometrics, Vol. 96, No. 2, 357–393.

Billingsley P., 1961, Statistical inference for Markov processes. The University of Chicago Press.

Blumen I., Kogan M. and Mac Carthy P.J., 1955, "The industrial mobility of labor as a probability process", Cornell Studies of Industrial and Labor Relations, Vol. 6, Ithaca N.Y., Cornell University Press.

Bonnal L., Fougère D. and Sérandon A., 1997, "Evaluating the impact of french employment policies on individual labour market histories", Review of Economic Studies, Vol. 64, No. 4, 683–713.

Burdett K., Kiefer N., Mortensen D.T. and Neuman G., 1984a, "Earnings, unemployment and the allocation of time over time", Review of Economic Studies, Vol. 51, No. 4, 559–578.

Burdett K., Kiefer N., Mortensen D.T. and Neuman G., 1984b, "Steady states as natural rates in a dynamic discrete choice model of labor supply", in: Studies in labor market dynamics, G.R. Neuman and N.C. Westergard-Nielsen (eds.), Berlin: Springer-Verlag, 74–97.

Burdett K., Kiefer N. and Sharma S., 1985, "Layoffs and duration dependence in a model of turnover", Journal of Econometrics, Vol. 28, No. 1, 51–69.

Butler J.S., Anderson K.H. and Burkhauser R.V., 1986, "Testing the relationship between work and health: a bivariate hazard model", Economics Letters, Vol. 20, 383–386.

Butler J.S., Anderson K.H. and Burkhauser R.V., 1989, "Work and health after retirement: a competing risks model with semiparametric unobserved heterogeneity", Review of Economics and Statistics, Vol. 70, No. 1, 46–53.

Card D. and Sullivan D., 1988, "Measuring the effect of subsidized training programs on movements in and out of employment", Econometrica, Vol. 56, 497–530.

Chesher A., and Lancaster T., 1983, "The estimation of models of labour market behavior", Review of Economic Studies, Vol. 50, No. 4, 609-624.

Chung K.L., 1967, Markov chains with stationary transition probabilities. Berlin: Springer - Verlag.

Doob J.L., 1953, Stochastic processes. New-York: Wiley.

Eberwein C., Ham J.C. and Lalonde R.J., 1997, "The impact of being offered and receiving classroom training on the employment histories of disadvantaged women: evidence from experimental data", The Review of Economic Studies, Vol. 64, No. 4, 655–682.

Eberwein C., Ham J.C. and Lalonde R.J., 2002, "Alternative methods of estimating program effects in event history mdoels", Labour Economics, Vol. 9, 249–278.

Eckstein Z. and Wolpin K.I., 1989, "The specification and estimation of dynamic stochastic discrete choice models", Journal of Human Resources, Vol. 24, No. 4, 562–598.

Elfving G., 1937, "Zur Theorie der Markoffschen Ketten", Acta Social Science Fenicae, series A2, No. 8, 1–17.

Flinn C.J. and Heckman J.J., 1982a, "New methods for analyzing structural models of labor force dynamics", Journal of econometrics, Vol. 18, 115–168.

Flinn C.J. and Heckman J.J., 1982b, "Models for the analysis of labor force dynamics", in: Advances in Econometrics, Vol. 1, R. Basmann and G. Rhodes (eds.), Greenwich, Conn.: JAI Press, 35–95.

Flinn C.J. and Heckman J.J., 1983a, "The likelihood function for the multistate-multiepisode model", in: Advances in econometrics, Vol. 2, R. Basmann and G. Rhodes (eds.), Greenwich, Conn.: JAI Press, 225–231.

Flinn C.J. and Heckman J.J., 1983b, "Are unemployment and out of the labor force behaviorally distinct labor force states?", Journal of labor economics, Vol. 1, No. 1, 28–42.

Fougère D. and Kamionka T., 2003, "Bayesian inference for the mover-stayer model in continuous-time with an application to labour market transition data", Journal of Applied Econometrics, Vol. 18, No. 6, 697–723.

Fougère D., Kamionka T. and Prieto A., 2005, "Stock-sampling bias in unemployment competing-risks duration models", mimeo, CREST-INSEE, Paris.

Frydman H., 1984, "Maximum likelihood estimation in the mover-stayer model", Journal of the American Statistical Association, Vol. 79, 632–638.

Geweke J., Marshall R.C. and Zarkin G.A., 1986a, "Exact inference for continuous time Markov chains". Review of Economic Studies, Vol. 53, 653–669.

Geweke J., Marshall R.C. and Zarkin G.A., 1986b, "Mobility indices in continuous-time Markov chains", Econometrica, Vol. 54, No. 6, 1407–1423.

Gilbert L., Kamionka T. and Lacroix G., 2001, "The Impact of Government-Sponsored Training Programs on the Labor Market Transitions of Disadvantaged Men", Crest Working Paper 2001-15, Paris.

Goodman L.A., 1961, "Statistical methods for the mover-stayer model", Journal of the American Statistical Association, Vol. 56, 841–868.

Gouriéroux C. and Monfort A., 1997, Simulation-based econometric methods, Oxford: Oxford University Press.

Gritz R.M., 1993, "The impact of training on the frequency and the duration of employment", Journal of Econometrics, Vol. 57, 21–51.

Ham J.C. and Lalonde R.J., 1996, "The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data on Training", Econometrica, Vol. 64, No. 1, 175–205.

Heckman J., 1990, "Alternative approaches to the evaluation of social programs ; Econometric and experimental methods", Barcelona Lecture, 6th World Congress of the Econometric Society, Barcelona, Spain.

Heckman J.J. and Borjas G.J., 1980, "Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence", Economica, Vol. 47, 247–283.

Heckman J.J., Lalonde R. and Smith J., 1999, "The economics and econometrics of active labor market policies", in: The Handbook of Labor Economics, Vol. 3A, O. Ashenfelter and D. Card (eds.), Amsterdam, North-Holland, 1865–2097.

Heckman J.J., and Singer B., 1984, "Econometric duration analysis", Journal of Econometrics, Vol. 24, 63–132.

Kalbfleisch J.D. and Lawless J.F., 1985, "The analysis of panel data under a Markov assumption", Journal of the American Statistical Association, Vol. 80, No. 392, 863–871.

Kamionka T., 1998, "Simulated maximum likelihood estimation in transition models", Econometrics Journal, Vol. 1, C129–153.

Kamionka T. and G. Lacroix, 2003, "Assessing the impact of non-response on the treatment effect in the Canadian self-sufficiency experiment", Crest working paper 2003-37, forthcoming in *Annales d'Economie et de Statistique*.

Kingman J.F.C., 1962, "The imbedding problem for finite Markov chains", Zeitschrift fur Wahrscheinlichkeitstheorie, No. 1, 14–24.

Klein J.P. and Moeschberger M.L., 2003, "Survival analysis: techniques for censored and truncated data", Second Edition, New York: Springer.

Lancaster T., 1990a, The econometric analysis of transition data. New York, USA, Cambridge University Press, 336p.

Lancaster T., 1990b, "A paradox in choice-based sampling", Working Paper 9017, Department of Economics, Brown University.

Lancaster T. and Imbens G.W., 1990, "Choice-based sampling of dynamic populations", in: Panel data and labor market studies, J. Hartog, G. Ridder and J. Theeuwes (eds.), Amsterdam, North-Holland, 21–44.

Lancaster T. and Imbens G.W., 1995, "Optimal stock/flow panels", Journal of Econometrics, Vol. 66, No. 1–2, 325–348.

Lindeboom M. and van den Berg G., 1994, "Heterogeneity in models for bivariate survival: the importance of the mixing distribution", Journal of the Royal Statistical Society, Series B, 56, 49–60.

Magnac T. and Robin J.M., 1994, "An econometric analysis of labour market transitions using discrete and tenure data", Labour Economics, Vol. 1, 327–346.

Magnac T., Robin J.M. and Visser M., 1995, "Analysing Incomplete Individual Employment Histories Using Indirect Inference", Journal of Applied Econometrics, Vol. 10, S153–S169.

Mealli F. and Pudney S., 1996, "Occupational pensions and job mobility in Britain: estimation of a random-effects competing risks model", Journal of Applied Econometrics, Vol. 11, 293–320.

Mealli F. and Pudney S., 2003, "Applying heterogeneous transition models in labor economics: the role of youth training in labour market transitions", in: Analysis of Survey Data, R. L. Chambers and C. J. Skinner (eds.), New-York, Wiley, 245–274.

Mortensen D., 1986, "Job search and labor market analysis", in: Handbook of Labor Economics, Vol. 2, O. Ashenfelter and R. Layard (eds), Amsterdam, North-Holland, 849–919.

Mortensen D.T. and Neuman G.R., 1984, "Choice or chance? A structural interpretation of individual labor market histories", in: Studies in labor market dynamics, edited by G.R. Neuman and N.C. Westergard-Nielsen, Berlin: Springer-Verlag, 98–131.

Olsen R., Smith D. and Farkas G., 1986, "Structural and reduced-form models of choice among alternatives in continuous time: youth employment under a guaranted jobs program", Econometrica, Vol. 54, 375–394.

Ondrich J., 1985, "The Initial Conditions Problem in Work History Data", Review of Economics and Statistics, Vol. 67, No. 3, 411–421.

Phillips P.C.B., 1973, "The problem of identification in finite parameter continuous time models", Journal of Econometrics, Vol. 1, No. 4, 351–362.

Ridder G., 1986, "An event history approach to the evaluation of training, recruitment and employment programmes", Journal of applied econometrics, Vol. 1, No. 2, 109–126.

Robert Ch. P. and G. Casella, 2002, Monte Carlo Statistical Methods. New York: Springer.
Runnenberg J.Th., 1962, "On Elfving's problem of imbedding a discrete-time Markov chain in a continuous-time one for finitely many states", Proceedings, Koninklijke Nederlandse Akademic van Wetenschappen, ser. A, Mathematical Sciences, Vol. 65, No. 5, 536–41.
Rust J., 1994, "Dynamic structural models; Problems and prospects: discrete decision processes", in C. Sims (ed): Advances in Econometrics, Sixth World Congress, Vol. II, Econometric Society Monographs, Cambridge: Cambridge University Press, 119–170.
Sampson M., 1990, "A Markov chain model for unskilled workers and the highly mobile", Journal of the American Statistical Association, Vol. 85, No. 409, 177–180.
Singer B., 1981, "Estimation of nonstationary Markov chains from panel data", Sociological Methodology, San Francisco. Samuel Leinhart-Jossey-Bass Publishers.
Singer B., 1982, "Aspects of non-stationarity", Journal of Econometrics, Vol. 18, 169–190.
Singer B. and Spilerman S., 1976a, "The representation of social processes by Markov models", American Journal of Sociology, Vol. 82, No. 1, 1–54.
Singer B. and Spilerman S., 1976b, "Some methodological issues in the analysis of longitudinal surveys", Annals of Economic and Social Measurement, Vol. 5, No. 4, 447-474.
Spilerman S., 1972, "Extensions of the mover-stayer model", American Journal of Sociology, Vol. 78, No. 3, 599–626.
Trivedi P.K. and Alexander J.N., 1989, "Re-employment probability and multiple spells: a partial-likelihood approach", Journal of business and economic statistics, Vol. 7, No. 3, 395–401.
Tuma N.B. and Robins P.K., 1980, "A dynamic model of employment behavior: an application to the Seattle and Denver Income Maintenance Experiments", Econometrica, Vol. 48, No. 4, 1031–1052.
Van den Berg G.J., 1997, "Association measures for durations in bivariate hazard rate models", Journal of Econometrics, Vol. 79, No. 2, 221–245.

Chapter 26
Software Review

Pierre Blanchard

26.1 Introduction

Since the previous edition of this book, econometric software used for estimating panel data models have been improved on three different scales. First, many packages allow now to estimate linear models on huge panel data sets made up of thousands of individuals with moderate execution times, even on a personal computer. Second, the estimation of linear dynamic panel data models is becoming more and more frequent. With several packages, it is now easy, i.e. with short programming, to estimate such models by instrumental variable method or by generalized method of moments. Lastly, the evolution has been the most significant in non linear model estimation on panel data. As shown in this book, an important literature on discrete choice and count models estimation with panel data, to name only a few, has been developed. Estimating such models is now an essential trend for applied econometricians and the need for appropriate econometric software is great.

Furthermore, a rapid glance to publications[1] frequently reviewing econometric software shows that the number of software used by econometricians (including their add-on modules and user application programs) has considerably increased. Even if we restrain ourselves to those which can be easily used with panel data, a great deal offer various econometric methods and provide different environments. Software for econometrics on panel data can broadly be classified into two main groups:

- First, we find general-purpose econometric packages, like LIMDEP, RATS, SAS, TSP... using command-driven languages, pull-down menus or both. They are

Pierre Blanchard
Erudite, Faculté de Sciences Economiques et de Gestion, Université Paris XII Val de Marne, 61 Av. du Général de Gaulle, 94010 Créteil Cédex, France, e-mail: blanchard@univ-paris12.fr

[1] For instance, *Computational Economics, Economic Journal*, *Journal of Applied Econometrics*, *Journal of Economic Surveys*...

L. Mátyás, P. Sevestre (eds.), *The Econometrics of Panel Data*,

easy to use and flexible to allow many, but not all, sorts of model estimation. For specific problems, they offer very often a matrix programming language.
- The second group consists of high-level matrix programming languages such as GAUSS, Ox... A common feature of these languages is that they are oriented towards intensive matrix manipulations. They require a good level of skill in programming, but they have the advantage of being largely flexible and, potentially at least, very fast. They are mainly used for hard to program estimation methods (non linear models), Monte-Carlo simulation...

Among the six software selected in the 1996's version of this chapter, five of them have made significant improvements and then will be reviewed here: four general-purpose econometric packages, LIMDEP, RATS, TSP, SAS and one high-level matrix programming language, GAUSS. Three others, not included in the previous edition, were added: two general-purpose econometric packages: EViews and Stata; one high-level matrix programming language: Ox.

Moreover, some very well known software are not included in this review, for instance, Shazam, SPSS... (general-purpose econometric packages), O-Matrix, R, S-Plus... (high-level matrix programming languages) and some specialized software (Frontier...). We did not review them because we need some basic econometric methods for panel data (within, between, FGLS, IV... estimators) which must be available without tedious programming (or with template programs and adds-ons).

In the first two sections of this chapter, we analyze the selected software belonging to each two groups enumerated above. Most of these software are regularly reviewed, so we discuss them mainly in terms of panel data management and relevant econometric methods they offer. We illustrate[2] their capabilities in estimating first, linear static and dynamic panel data models and, second (when possible), several non linear models (a random effects model and a random effects probit model estimated by maximum likelihood method).

As emphasized by McCullough and Vinod (1999), numerical accuracy is an important, but often neglected, characteristic of econometric software. Section 26.3 is devoted to this problem and to performance evaluation.

26.2 General-Purpose Econometric Packages

26.2.1 EViews (v. 5.1)

EViews (Econometric Views) is mainly used for graphical and statistical analysis, time series estimation and model simulation (see Roberts, 1995, Holly and Turner, 1998, Sparks, 1997 and Noble, 1999). Thanks to version 4, it was already possible to estimate some panel data models with the `pool` object, which is useful

[2] All the programs, data and web links used in this chapter can be found on a companion web site of this volume (`http://www.univ-paris12.fr/epd`).

Table 26.1 EViews code #1 (grunfeld_eviews.prg)

```
cd c:\epd\data\eviews                              ' default directory
wfcreate(wf="grunfeld") a 1935 1954 10 ' creating the workfile
read(t=xls,a2) "grunfeld.xls" 11        ' loading the excel file
```

when dealing with a panel dataset with small N and large T. The last release[3] offers now extended panel data management capabilities and estimation features for balanced and unbalanced panel, particularly when the panel datasets contain many individuals. One of the interesting features of EViews is that this software is designed to be used mainly by pull-down menu. A program mode (with also a command line interface) is also available, and, for space limitation reasons, we limit our presentation to the batch mode, which is not necessarily the simplest one. Another particular aspect of this software is the concept of workfile which is central to EViews' working. A workfile is a container for EViews objects (equations, graphs, matrices, series...) and each workfile may be formed of one or more workfile pages (each containing specific objects). EViews can manage series of maximum 4 million observations for each, and the total number of observations is only limited by available RAM size.[4] When a user works on a panel data set, its workfile will contain several series which are associated with an individual identifier and a date identifier. Entering panel data in the workfile depends on the nature of the panel data set.

If the panel is a balanced one and if the basic dataset is, for instance, in an Excel format, the following code[5] (see Table 26.1) will execute this task for the Grunfeld data ($N = 10$, $T = 20$).

In the `wfcreate` instruction, `a` stands for annual data (beginning in 1935 and ending in 1954, but other periodicities instead of annual may be used) and `10` indicates than the panel contains 10 individuals. The `read` instruction can import ASCII, Lotus and Excel files. For reading more file formats (Stata, SAS transport, HTML, GAUSS, GiveWin, RATS, TSP...), in addition to the previous ones, one can use the `wfopen` instruction (see Table 26.2).

If the panel is unbalanced, you must follow three steps. First, you have to create an unstructured workfile with the `wfcreate` instruction. Second, read the external data file (or even an EViews file) with a `read` or `wfopen` instruction. Third, it is necessary to structure the workfile as an unbalanced panel with the `pagestruct` instruction applied to an individual identifier (say `ident`) and to a date identifier

[3] EViews 6 is now available.

[4] For instance, with 256 Mb RAM, a workfile may contain more than 1900 series with 100,000 observations each.

[5] In this chapter, sofware commands, options and Web links are written in `Typewriter font`. For all softwares, except EViews, Ox and Stata, the first letter of a command is written in uppercase; file names and (user) variable names are written in lowercase. For Stata, commands, file names and variables names are always in lowercase. For Ox, names are case sensitive. We follow the EViews documentation in writing all in lower cases.

Table 26.2 EViews code #2 (ab_eviews.prg)

```
cd c:\epd\data\eviews                              ' default directory
wfcreate(wf="ab") u 1 1031       ' creating the undated structure
wfopen "c:\epd\data\excel\ab.xls"        ' loading the excel file
pagestruct ident @date(year)       ' creating the panel structure
```

(say `year`). These steps are illustrated by the following piece of code using data from Arellano and Bond, 1991 (hereafter AB), saved initially in an Excel file.

Of course, if the data are already saved in a panel data workfile, `wfopen "c:\epd\data\excel\ab.xls"` would be sufficient. If the workfile has not a panel structure, add only a `pagestruct` instruction (you will need an individual identifier and a date identifier) after the instruction `wfopen`.

Moreover, a `sort` instruction with several sort keys is available and merging workfiles is also possible by using links facilities (easy to do by menus). Alphanumeric series and several date formats are recognized in a workfile, but EViews considers alphanumeric series in an Excel or an ASCII files as missing values (coded by `na`). A complete programming language may be used for complex tasks.

Once, your workfile is ready, graphics, descriptive statistics and estimation on panel data estimation are quite straightforward. First, EViews offers several useful tools (statistics, graphics, test for equality of means and variances between groups...) as shown by the following program reproduce in Table 26.3.

This code illustrates another EViews characteristic: the object concept. For instance, the previous program defines a group object (named `g1` referring to the variables `year` and `gi`). Then, we can apply some view (a graph, some tests for instance) or some procedures (e.g. an equation estimation) to this object. Static panel data estimation is easily carried out by few instructions. Nevertheless, there is one exception: the between estimator which requires a great deal of programming (this cannot be done by menus). We do not report how to obtain between estimation due to space limitations (see `grunfeld_eviews.prg`).

The `ls` instruction (see Table 26.4) offers several options which are very useful for panel data estimation, particularly several ways to apply robust methods for computing the coefficient standard errors (White, SUR...). One-way and two-way specifications are also supported as well as different ways to estimate the component variances in random effects models (Swamy-Arora, Wallace-Hussain, Wansbeek-Kapteyn).

Table 26.3 EViews code #3 (grunfeld_eviews.prg)

```
vc.statby ident          ' descriptive statistics by individual
group g1 year gi               ' defining a group of variables
g1.xy(panel=individual)                    ' individual graphs
gi.testby(mean) ident                                 ' anova
```

Table 26.4 EViews code #4 (grunfeld_eviews.prg)

```
equation ols.ls gi c vf vc                                    ' ols
equation within.ls(cx=f) gi vf vc c                         ' within
equation ec.ls(cx=r) gi vf vc c                               ' fgls
ec.ranhaus                                                'Hausman test
```

Dynamic panel data estimations[6] are equally easy to obtain by the use of menu (Proc and by choosing the GMM/DPD method) or by the Dynamic Panel Wizard (a succession of 6 dialog windows). This can be also done by programming as indicated in Table 26.5 which can reproduce the AB results (Tables 4 and 5, pp. 290, 292). We can also apply the Anderson–Hsiao estimator (hereafter AH, see Anderson and Hsiao, 1981). Note that the underscore ("_") is the continuation line character.

This program requires several explanations.

1. First, note the use (new to EViews 5) of replacement variables (called sometimes macro variables) quite useful for defining variable lists used repetitively. For instance, `%inst = "w w(-1) k k(-1) k(-2) ys ys(-1) ys(-2)"` defines a list of instruments which can be used when {`%inst`} is referred to in the program. Remark also that there is only one instruction in each line. A long instruction may be set in several lines with the _ (underscore) character.

Table 26.5 EViews code #5 (ab_eviews.prg)

```
' define replacement variables
%rhs = "n(-1) n(-2) w w(-1) k k(-1) k(-2) ys ys(-1) ys(-2)"
%model = "n " + %rhs
%inst = "w w(-1) k k(-1) k(-2) ys ys(-1) ys(-2)"
' (5g) OLS with White robust se
equation ols.gmm(per=f,gmm=perwhite) {%model} @ {%rhs}
' (5h) LSDV with White robust se
equation within.ls(cx=f,per=f,cov=perwhite,nodf) {%model}
' (5e) AH with n(-3) in diff.,White robust se.
equation ahd.gmm(cx=fd,per=f,levelper,cov=perwhite,nodf)    _
  {%model} @ n(-2) n(-3) {%inst}
' (5f) AH with n(-3) in lev., White Period Robust se.
equation ahl.gmm(cx=fd,per=f,levelper,      _
     gmm=ident,cov=perwhite,nodf)             _
     {%model} @ n(-2) @lev(n(-3)) {%inst}
' (4a1) AB first-diff., period dum. in lev., 1st step robust.
equation aba1.gmm(cx=fd,per=f,levelper,cov=perwhite,nodf)    _
    {%model} @ @dyn(n) {%inst}
' (4a2) - AB first-diff., period dum. in lev., 2nd step robust
equation aba2.gmm(cx=fd,per=f,levelper,gmm=perwhite)          _
      {%model} @ @dyn(n) {%inst}
```

[6] Instrumental variable estimation can be obtained with the `tsls` instruction.

2. EViews propose several functions that are useful for panel data. `@expand(ident)` owes to create automatically individual dummy variables (one for each individual in the panel). There are also instructions for creating trend variables (`@trend, @trendc...`).
3. Option `cx=` requires fixed effects estimation when `cx=f`, first difference estimation when `cx=fd` and orthogonal deviation when `cx=od`. By default, no transformation is done. When `per=f`, time dummies are included, and if `levelper` is specified, time dummies are in levels even if `cx=fd` or `cx=od`.
4. `gmm` and `cov` Options define GMM weighting (identity, White...) and coefficient covariance computation method (White, SUR, ordinary...).
5. The special instruction `@dyn(n,-2,-5)` permits to include lags of `n` from 2 to 5 as instruments as suggested by Arellano and Bond, (1991). With `@dyn(n)`, EViews will incorporate all the possible lags.

There are other EViews capabilities that have to retain our attention. EViews 5 is the first software which offers panel data unit root tests without any programming. In recent years, there has been growing interest in the use of macroeconomic panel data (sets of countries, regions or industries -large N- on many periods -large T-). It is frequently advocated that panel data unit root tests have higher power than those done on time series data. With EViews, several tests are available (provided with many options): (1) Levin, Lin and Chu, (2) Breitung, (3) Im, Pesaran and Shin, (4) Fisher-type using ADF (5) Fisher-type using PP and (6) Hadri tests. The availability of these tests associated to many estimation methods on pooled time-series cross-section data (T large, N small) will certainly be useful for panel data macroeconomists.

Up to now, non linear model estimation (binary probit/logit, ordered models, censored models, count models...) are limited to cross section data. Nevertheless, EViews offers a specific procedure (called the Log Likelihood Object) to estimate other non linear models by the maximum likelihood method (see `chronoml_eviews.prg` used in Sect. 26.4.1). Yet, it seems difficult to adapt it to non linear panel data estimation (random effects probit model, by instance) when it is necessary to evaluate separately the log likelihood by individuals on T_i periods.

The on-line help and the paper documentation (2 volumes, more than one hundred pages on pooled and panel data management and estimation) are quite good. The web site contains mainly commercial information. A EViews group discussion via E-mail or via Usenet is not available. It would be useful that, in the future, more program examples on panel data will be downloadable.

26.2.2 LIMDEP (v. 8) with NLOGIT (v. 3)

LIMDEP[7] (LIMited DEPendent variable models) was initially designed, as its name indicates, in order to estimate models having limited or qualitative dependant

[7] LIMDEP 9 and NLOGIT 4 are available.

variables. Successive releases have improved its coverage of econometric methods in many domains.

As noted by McKenzie and Takaoka (2003), LIMDEP Version 8 (with 3.0 for Nlogit) "represents a significant expansion of the estimation techniques for panel data models". For some discrete choice models estimation on panel data, you will need Nlogit, available at a substantial extra cost. In all cases, very few programming is required.

LIMDEP requires a microcomputer running under Windows 95 or a later version and works either in batch mode or by menu. LIMDEP uses four key windows: the output window (which reports log and results), the project window, the command window and the data editor window. As our focus is concentrated only on batch mode, we let aside the use of menu and use mainly the command window. On this point, one of LIMDEP's interesting feature is that it produces a trace file (`trace.lim`) which gives a complete trace of the LIMDEP's session (list of commands obtained when using the mode menu, for instance). Further details on LIMDEP's general features may be founded in Fry (1996), Heywood and Rebelo (1996), Wright (1996), McCullough (1999b), McKenzie (2000) and McKenzie and Takaoka (2003).

As it is simply impossible to give a complete list of the LIMDEP capabilities in panel data field, we sketch only the most important. Note that, generally speaking, little programming is required.

As we focus on the batch mode, we only need to describe the panel data file and the file (`.lim`) containing LIMDEP's code .

Firstly, LIMDEP may manage large (balanced or unbalanced) panel datasets with a limit[8] of 900 variables (see the `rows` instruction). LIMDEP may read various file formats, mainly ASCII, XLS, binary... A panel data is supposed to be organized by individuals; if not, a `Sort` instruction with one key is available for re-ordering the panel if necessary (you may use too the `Reorder` options of `Regress` command). Merging panel datasets is not possible only in the case when you have to merge a panel dataset containing invariant variables with a usual panel dataset (in `Read` instruction). Obviously, you need a common individual identifier. So, that is why one needs (unless your data are balanced) a panel dataset that contains an individual identifier (say `ident`) which can be of any numerical values. Note also that missing values are managed in a very particular way by LIMDEP. Missing values are coded with -999. Whether these observations are eliminated for computations or not depends on the command being used. For panel data estimation, LIMDEP will exclude all the rows for an individual if one observation is missing. Notice that, with panel data, the `Create` instruction like `Create; x1 = x[-1] $` does not take into account the switching from an individual to another one. In this case, you can use the `Reject` instruction to eliminate some observations, but this may contradict the `Pds` specification (see below, but don't use the `Skip` instruction). So, users have to be very careful, in particular with dynamic models.

[8] See Limdep's documentation for a complete list of program limits. For instance, with the fixed effects binary logit model, 100 periods for one individual is a maximum.

Table 26.6 Limdep code #1 (grunfeld.lim)

```
Reset $
Read ; File = C:\epd\data\excel\grunfeld.xls; Format=Xls ; Names $
Title ; ols, within and fgls estimations $
Regress ; Lhs = gi ; Rhs = One, vf, vc ; Pds=20 ; Panel $
Title ; between estimation $
Regress ; Lhs = gi ; Rhs = One, vf, vc ; Pds=20 ; Means ; Panel $
```

Secondly, the definition of the panel data structure is set up when specifying the estimation instructions, and not when reading the data file. In Table 26.6, we present a simple example on Grunfeld's data.

This short program suggests several remarks:

- In batch mode, you have to be careful in typing commands which use $ and ; symbols in a different way from other software (TSP, SAS...). Commands end with a `$` and `;` is the symbol for separating options (if any) in a command. Comments are included with the `? comment ;` or with the `/* comment */` syntax. In the same way, LIMDEP's programming language may be different from others in the way it treats loop's instructions, procedures etc.
- As our panel is balanced, we specify `Pds = 20`, because the panel has 20 periods by individual (a group for LIMDEP). If the panel was unbalanced, you have to write `Pds = ni` where `ni` is a variable which gives, each individual, the number of period observations (for instance $4,4,4,4,2,2,3,3,3$). Our model being a linear one,[9] it would be possible to replace the `Pds` specification by `Str = ident` where `ident` is a numerical variable containing a unique numeric identifier code. It is possible to create a `ni` (and a new `ident identifier`) variable by the following code: `Regress ; Lhs = One; Rhs = One; Str = ident ; Panel $`. This will create two new variables: `_groupti` which contains the number of observations by individual, and `_stratum` containing a new individual identifier equal to $1,2,\ldots,N$. Another useful instruction (not shown above) `Expand,` may transform a categorical variable into a set of dummy variables.
- Our `Panel` option requires an estimation for the panel data, in our example, a one-way fixed or random effects linear models. Through `Means`, `Fixed Effects` or `Random Effects` options, it is possible to obtain only specific results. Note also, that the output (not shown) is very clear and detailed; many useful tests are automatically reported (Hausman, F, LR and LM tests). Other tests may be programmed with LIMDEP's programming language. For a two-way specification, just specify `Period = timev` where `timev` is a variable name containing integers like $1,2,\ldots,T_i$ but not like `timev` = 1981, 1982,...,1999.

[9] For non linear models estimation on panel data, the `STR` specification is not possible, you must provide `Pds = varname` or `Pds = number`.

- Many other possibilities are available (with specific instructions or program templates): Restricted estimation, White robust estimation, random effects model estimation with autocorrelated error structure, estimation by 2SLS (for fixed effects model), the Hausman-Taylor (hereafter HT) estimator for random effects model, estimation of Hildreth-Houck and Swamy's random coefficients models etc.

For dynamic linear panel data models, Arellano–Bond-Bover GMM's estimator is available with LIMDEP, but it correctly works only with LIMDEP 9 (see the program called `ab.lim`, not reproduced here).

As far as non linear models on panel data are concerned, the range of LIMDEP capabilities is huge. With LIMDEP, you can estimate probit/logit, truncation, ordered probit/logit, poisson, negative binomial, bivariate probit, sample selection and stochastic frontier models with either fixed or random effects. With the Nlogit package, multinomial logit, heteroscedastic extreme value, random parameters logit, nested logit, latent class, multinomial multiperiod probit models among others may also be estimated. We will study only two illustrations. First, in Table 26.7, we estimate (on Grunfeld's data) a one-way linear random effects model by ML.

This example shows how we can use the `Gxbr` instructions for computing individual means, and the deviations to individual means. It indicates also that it is easy to write the log-likelihood for an observation. Note that it is possible to supply (not shown here) first (but not second) derivatives when speed and convergence problems are important. If the log likelihood has to be computed in using separately the T_i log-likelihood for an individual, it will be more complicated.

Fortunately, for many situations, LIMDEP does it automatically. Very few programming is then required for estimating mixed, binary, multinomial probit/logit... models. For instance, in Table 26.8, we estimate on Keane and Wolpin's data (hereafter KW, see Keane and Wolpin, 1997) a random effects binary probit model.

Table 26.7 Limdep code #2 (grunfeld.lim)

```
? creating the individual means (used by ml)
Matrix ; mgib = Gxbr(gi,ident) ; mvfb = Gxbr(vf,ident) ;
   mvcb = Gxbr(vc,ident) $
Create ; gib = Mgib(ident) ; vfb = Mvfb(ident) ;
   vcb = Mvcb(ident) $
? creating deviations to individual means
Create ; giw = gi - gib ; vfw = vf - vfb ; vcw = vc - vcb $
Title ; ML estimation of error components model $
Calc ; tt = 20 ; nn = 10 ; nt = 200 $
Maximize ; Start = 1.0 , 1.0 , 1.0 , 1.0 , 1.0 ;
           Labels = b0 , b1 , b2, s2u , s2a ; Alg = Bfgs ;
 Fcn = -Log(2*3.14116)/2-Log(s2u)*(tt-1)/(2*tt)
       -Log(s2u+tt*s2a)/(2*tt)-(1/(2*s2u))*((giw - b1*vfw -
   b2*vcw)^2)
       -(1/(2*(s2u+tt*s2a)))*((gib - b0 - b1*vfb -
   b2*vcb)^2) $
```

Table 26.8 Limdep code #3 (discrete_choice.lim)

```
Reset $
? NB: ident (1 to 1838), year (81 to 87)
? NB : choice 1= school, 2=home, 3=wcollar, 4=bcollar, 5=serv
? Reading an Excel data file
Read ; File = c:\epd\data\excel\kw.xls;Format=XLS;Names $
? creating a variable named _groupti given Ti
Regress ; Lhs = One ; Rhs = One ; Str = ident ; Panel $
? Creating the 0/1 variable for binomial models
Create ; status = choice > 2 $
Title ; one way random effects probit model $
Probit;Lhs=status;Rhs=One,educ,exper,expersq,black;
   Pds=_groupti ; Random Effects; Maxit = 500 $
Stop $
```

The documentation is good, large (may be too large, one Reference Guide and two Econometric Modeling Guides, plus a separate documentation for NLOGIT, all with many examples) and sometimes we encounter some repetitions (see, for instance, § R6-6 and § R16-24). Nevertheless, even if you have not planned to use LIMDEP, you ought to read, at least, the two chapters entirely devoted to panel data models (chapter E8, Linear Models for Panel Data and E14, Panel Data for Binary Choice). Indeed, they contain very interesting information concerning these models, in particular, many illuminating technical details and very judicious notes on practical aspects of econometric modeling with panel data. The LIMDEP's listserv is of equal interest.

26.2.3 RATS (v. 6)

Even if RATS (Regression Analysis of Time Series) is mainly designed for time series analysis, it also provides a few special instructions to estimate linear models on panel data.

One interesting RATS' feature is that it deals explicitly with the panel data structure. This can be done with the `Calendar` and `Allocate` instructions[10] which define the data structure. For instance, if you use the Grunfeld's data (a balanced panel dataset), you may write, at the beginning of your RATS program, the following code:[11]

```
Calendar(Panelobs=20) 1935 1 1
Allocate 10//1954:1
```

[10] `Allocate` is now optional with RATS 6.

[11] A Wizard environment is also available for some simple tasks.

This panel data set have 10 individuals (firms), observed between 1935 and 1954 (20 years). The observation for a variable, say `y`, for individual n°3 observed in 1950 is referenced like `y(3//1950:1)`. Note that, for RATS, a panel data set is always organized by individual (all time periods for the first individual, followed by all time period for the second individual,...). Moreover, the periodicity of the panel may be annual, quarterly, biannual, monthly, weekly... In adjusting the `Calendar` and `Allocate`, you could extract easily a sub sample from the file too.

The panel doesn't need to be balanced, as it is the case for AB data. Some caution is yet required: if you are reading a non RATS data file (Excel for instance), your file must be "balanced" in a special sense, i.e. each individual must be observed during the same period, but when an individual is not observed in a given year, variables (but not the date's entry) have to be set to missing values (`%NA` for RATS). You can balance it manually or in using the `Pform` instruction (see `ab_rats.prg` for an illustration). This instruction may reorganize unbalanced panel data series (an index series for the individuals and for the time periods are then required); RATS inserts missing values for all variables when a year for an individual is missing. When the operation is realized, you can save your "balanced panel dataset" in a RATS data file. It is then possible, for instance, that an individual begins in 1976 and stops in 1986 and another one begins in 1979 and stops in 1983.

Suppose now that your panel data set is well defined. Thanks to the `Calendar` and `Allocate` instructions, creating new series becomes very simple. Particularly, if you create a lagged series by `SET yl1 / = y{1}`, the series `yl1` contains a missing value in the first year of each individual. This same property applies to differenced and led series. Moreover, RATS documentation provides simple code for creating trend series, individual and time dummies using specific panel function `%trend(t)`, `%indiv(t)` and `%period(t)`. When your variables (the series for RATS) are correctly defined, you have several ways for estimating a panel data model.

RATS offers various instructions to estimate a static linear model with random or fixed effects (one-way or two-way).

To estimate a (one-way) random and fixed effects model on the Grunfeld's data, the most simple way is to use the `Pregress` instruction, see Table 26.9.

Unbalanced panels are correctly taken into account in the computation of θ_i.

Another way that aims at estimating a (one-way) fixed or random effects model consists in using, first the `Panel` instruction to transform the series in taking deviation to the individual means, and then to estimate the transformed model by OLS. As an example, in order to apply the within estimator, we can transform the original series in deviation to the individual means by instructions reported in Table 26.10 which create $giw_{it} = gi_{it} - gi_{i.}, \ldots$ and then, estimate the transformed model by OLS.

Note that `(Dfc=10)` is necessary for correcting the number of degrees of freedom that are lost by substracting out the 140 (we suppose $N = 140$) individual means. The `Pregress` instruction do not have a constant term.

We can proceed in a similar way to estimate a model by between or FGLS estimator. Moreover, a random effects model can be also estimated in using first, the instruction `Pstats` applied to the OLS residuals to obtain estimations of a σ_u^2 and

Table 26.9 RATS code #1 (grunfeld_rats.prg)

```
End 1
Calendar(Panelobs=20) 1935 1 1
Allocate 10//1954:1
* Reading the xls data file
Open Data C:\epd\data\excel\grunfeld.xls
Data(Unit=Data,Format=Xls,Org=Obs) / ident year gi vf vc
Close Data
* within
Pregress(Effects=Indiv,Method=Fixed) gi
# vf vc
* fgls
Pregress(Effects=Indiv,Method=Random) gi
# Constant vf vc
```

σ_ε^2, second, in computing θ_i by usual formula, third in transforming the original series with the `Panel` instruction under the form $gire_{it} = gi_{it} - \theta_i gi_{i.}, \ldots$ and finally in applying OLS on transformed series.

From the point of view of the panel data model estimation process, RATS is flexible and can deal with large panel datasets thanks to its use of virtual memory. Nevertheless, the output regression is incomplete, only an F test for the presence of individual (or time) specific effects and a LR test for equal variance of perturbations across individuals (with the `Pstats` instruction) are reported. Neither Hausman test, nor other tests for the presence of individual (or time) specific effects (two sided and one sided LM, ...) are available without programming. Of course, it is always possible to program them with RATS in using its matrix language or with users's procedures (if any).

On the Estima Web site, there are several programs that can be used in estimating dynamic linear models on panel data by GMM. However some programming efforts are required as, for instance, in the design of the instruments matrix used by AB estimator (see `ab_rats.prg`). Nevertheless, in simpler cases, for instance, estimating a dynamic panel data model by AH method is easy to realize. It is achieved by the code given in Table 26.11.

There are no specialized instructions for estimating non linear models on panel data (probit random effects model,...). But, maximum likelihood estimation may be done using the `Maximize` instruction. For instance, with RATS, it is easy to estimate a random effects model by ML. One needs only to write the likelihood for an observation using individual means and deviation to individual means created

Table 26.10 RATS code #2 (grunfeld_rats.prg)

```
Panel(Entry=1.0,Indiv=-1.0) gi / giw
Panel(Entry=1.0,Indiv=-1.0) vf / vfw
Panel(Entry=1.0,Indiv=-1.0) vc / vcw
Linreg(Dfc=10) giw
# vfw vcw
```

Table 26.11 RATS code #3 (ab_rats.prg)

```
Calendar(Panelobs=9) 1976 1 1
Allocate 140//1984:1
* reading the data file and variable creation not reproduced
* computing variables in difference
Diff n / dn ; Diff w / dw ; Diff k / dk ; Diff ys / dys
* estimation by Anderson-Hsiao IV, cf. AB (1991), pp. 292 col e
Instruments dn{2 3} dw{0 1} dk{0 1 2} dys{0 1 2} dumyear
Linreg(Optimalweights,Instruments,Robusterrors,Lags=T-1) dn
# dn{1 2} dw{0 1} dk{0 1 2} dys{0 1 2} dumyear
```

by the `Panel` instruction. Remember that RATS requires that you write the log likelihood for an individual i at time t, which is done by the `Frml` instructions reproduced in Table 26.12.

If your log likelihood involves computation of individual log likelihood by product of time period observations (for instance as in a random effects probit model), this will be more difficult to realize. Moreover, as it seems impossible to define first and second derivatives, it will be difficult to achieve convergence in some complicated cases.

More generally (cf. McCullough (1997) and Heywood and Rebelo (1996)), RATS works through an interactive mode as well as in batch mode and is available for many platforms (Windows, Macintosh, Unix, VMS...). RATS has a simple but good interface (basically, a program window and an output/log window). It can read and write numerous file formats (ASCII, XLS, DBF[12]...) and also has some databank management functions. Nevertheless, a merge instruction is not offered and the sort instruction (`Order`) admits only one sort key, which is not always adequate for some panel data management. Another deficiency is the lack of cross-tabulation instruction.

The Estima Web site contains a wide range of useful information, in particular for panel data, some RATS programs or procedures for (Pedroni) unit root and

Table 26.12 RATS code #4 (grunfeld_rats.prg)

```
Nonlin b1 b2 b3 s2u s2b
Compute b1=b2=b3=s2u=s2b=1
Frml resww = giw - b2*vfw - b3*vcw
Frml resbb = gib -b1 - b2*vfb - b3*vcb
Frml logl = -0.5*Log(2*%PI) -Log(s2u)*(tt-1)/(2*tt) $
        -Log(s2u+tt*s2b)/(2*tt) -(1/(2*s2u))*(resww**2) $
        -(1/(2*(s2u+tt*s2b)))*(resbb**2)
Maximize(Iter=2000,Notrace,Print) logl /
```

[12] This format is quite useful for file conversion because DBF file are not restricted on the number of records and of variables, contrarly to Excel which is limited to 65536 lines and 256 columns.

cointegration testing with panel data, estimation of spatial correlation consistent covariance matrix from panel data, estimation of Swamy's random coefficient model.

In addition, you can download some panel data example programs with their data sets to reproduce the results presented, for instance, in Wooldridge (2002).

The documentation is good (many examples with data & programs). Nevertheless, in some cases, computational methods are not fully described, for instance: how does the `Pstats` instruction work? how are obtained the estimations of a σ_u^2, σ_ε^2 and then θ_i with the `Pregress` instruction? This may be not a problem due to the very effective working of the RATS discussion list.

26.2.4 SAS (v. 9.1)

SAS is probably the best known of all statistical software packages and is available virtually on all platforms except on Macintosh computers. Microcomputer and mainframe versions have exactly the same features with a homogeneous user's interface (mainly three windows: program, log and output ones). Working in interactive and batch modes, SAS covers a wide range of statistical and econometric methods available in several modules which can be rented separately (SAS/Base, SAS/Stat, SAS/ETS, SAS/Graph...).

An SAS program consists usually in series of Data steps which create SAS tables on which we apply Procedure steps to carry out statistical, graphical and other types of analysis. SAS uses virtual memory and offers a powerful macro-language (SAS/Macro included in SAS/Base) and a good matrix programming language (SAS/IML) as a separate module. Even if some improvements had been reached with the last release, SAS is very often criticized for its outdated features which do not support many econometric estimation methods and econometric testing. This is particularly true for panel data estimation.

SAS provides very few specific procedures for static linear model estimation on panel data. The most useful one is `Proc Tscsreg` which is illustrated in Table 26.13 applied to Grunfeld's data.

Several points stand out in `Proc Tscsreg`. First, the panel dataset may be either balanced or unbalanced using the `Id` instruction for balanced (or unbalanced data or `Cs=` and `Ts=` options for balanced data). Second, two-way model estimations are possible in specifying `Fixedtwo` or `Rantwo` options and some variants are available (first-order autoregressive structure for the error term...). Third, this procedure reports an Hausman test and an F test for no fixed effects. Finally, with a great number of individuals (> 5000 with the configuration used in this review), a virtual memory problem may appear when using the within estimator (probably due to the computation of fixed effects without using Frisch-Waugh method). In this case, it is better to use `Proc Gml` with the `Absorb` (using Frisch-Waugh method) instruction as shown in Table 26.13.

Nevertheless, FGLS estimation is always very slow with numerous individuals (with `Proc Tscsreg`) due to the preliminary estimation of the fixed effects

Table 26.13 SAS code # 1 (grunfeld.sas)

```
Libname in 'c:\epd\data\sas' ; Run ;

Proc Tscsreg Data = in.grunfeld ;
    Id ident year ; Model gi = vf vc / Fixone ; Title within ;
Run ;

Proc Tscsreg Data = in.grunfeld ;
   Id ident year ; Model gi = vf vc / Ranone ; TITLE fgls;Run ;

Proc Glm Data = in.grunfeld ; Model gi = vf vc ; Absorb ident ;
Title within with Proc Glm ; Run ;

Proc Mixed Data = in.grunfeld ; Class ident ;
   Model gi = vf vc / S ; Random ident / S ; Method = Ml ;
Title ML estimation of RE model ; Run ;
```

models used for computing σ_ε^2. So, very often, users prefer first, to transform the variables in computing individual means, differences and quasi-differences to individual means and second, to estimate the transformed models by OLS obtaining between, within or FGLS estimates. The resulting program (not reported here because of space limitation) is longer but faster in executing on large panel datasets.

For maximum likelihood estimation of the random effects (linear) model, a simple solution (very time consuming with numerous individuals) may be implemented with `Proc Mixed`[13] (see Table 26.13).

There is a new (experimental) procedure for estimating dynamic linear model on panel data, `Proc Panel`, but, to our knowledge, it does not correctly work for dynamic models estimated by GMM. Note however that `Proc Panel` allows to create new variables (lags for instance) in the body of the procedure. The `Proc Model` instruction which may estimate models by IV and GMM works only on time series or cross section data. It is easy to implement AH methods, but `Proc Model` cannot reproduce standard errors robust to general cross-section and time series heteroscedasticity as in Arellano and Bond (1991). This may be done with a SAS/IML program, so more complex programming is required (this is done in `ab.sas`).

The situation is less controversial for non linear model estimation on panel data. Even if many procedures which implement Logit/Probit estimation are mainly designed for cross-section data, others work with panel data.[14] SAS may estimate fixed effects logit models with conditional maximum likelihood, generalized estimating equations (GEE with `Proc Genmod`), random effects logit and probit models on

[13] Note that SAS offers an other procedure for ML estimation, `Proc Nlp` which is part of SAS/OR module. `Proc Mixed` must be prefered to `Proc Nlin` which can estimate some models by ML (with the `model.like = logl ;` and `_loss_ = -logl ;` instructions). See `grunfeld.sas` for an illustration.

[14] See Allison (1999) for a detailed presentation.

panel data. As a SAS program spends a great part of its time reading and writing data on a hard disk, time execution for some non linear models may become extremely long, and sometimes, problems like insufficient memory occur. The instructions reported in Table 26.14 illustrate some SAS capabilities in discrete choice model estimation on KW data (variables creation not reported).

Another difficulty can be relevant with SAS. Indeed, there are many procedures to estimate discrete choice models (`Proc Logistic`,[15] `Proc Mixed`, `Proc Catmod`, `Proc Nlmixed`, `Proc Glm`, `Proc Probit`...) each with many different options. Therefore, it may be difficult to determine which one corresponds more to the purpose. Sometimes, the same results may be achieved by several procedures. Two procedures were recently introduced: `Proc Mdc` to estimate mixed logit models and `Proc Qlim` for (mainly) multinomial logit and tobit models, but only on cross-section data.

A crucial point must be here taken into consideration. For some procedures (`Proc Model`, `Proc Nlmixed`, `Proc Nlin`...), it is possible to insert programming statements (Do-loop, If expressions, arrays manipulation, macro-instructions...), so that, the user may adapt, for example, the computation of his log likelihood to a special problem. Yet, this way of programming may be difficult.

Table 26.14 SAS code # 3 (discrete_choice.sas)

```
/* conditional fixed effects logit model */
Proc Logistic Data = temp ; Model choice = educ exper expersq ;
  Strata ident ; Run ;

/* random effects logit model */
Proc Nlmixed data = in.kw ;
  Parms b0=1 b1=0 b2=0 b3=0 b4=0 s2u=1 ; pi = Constant('PI');
  eta = b0 + b1*educ + b2*exper + b3*expersq +b4*black+ u ;
  expeta = exp(-eta) ; p = 1/(1+expeta) ;
  Model choice ~Binary(p) ; Random u ~Normal(0,s2u) Subject =
   ident ;
  Estimate "rho" s2u/(s2u+(pi**2)/3) ; Run ;

/* random effects probit model */
Proc Nlmixed data = in.kw ;
  Parms b0=1 b1=0 b2=0 b3=0 b4=0 s2u=1 ;
  eta = b0 + b1*educ + b2*exper + b3*expersq + b4*black + u ;
  IF (choice = 1) Then p = Probnorm(eta) ; Else p = 1 -
  Probnorm(eta) ;
  ll = log(p) ; Model status ~General(ll) ;
  Random u ~Normal(0,s2u) Subject = ident ; Estimate
   "rho" s2u/(s2u+1) ;
Run ;
```

[15] `Proc Logistic` models Prob(y=0). Using the `descending` option allows to model Prob(y=1) instead.

However, it is hard to believe today that the scope of panel instructions would improve in the future. Indeed, the efforts aimed at developing SAS are massively concentrated upon the modernization of other parts of the software (web integration, data mining, data warehousing...). Here again, because of the vast community of SAS users, it is often easy to find SAS programs (with macro instructions and/or IML) for panel data estimation. Nevertheless, in the future, it is possible that the new version of IML, called SAS/IML Workshop (yet available under a test version) changes considerably this situation. Indeed, in an IML procedure, DATA and PROC steps cannot be integrated[16] into the IML program. IML Workshop enables you to utilize DATA and PROC steps (and external C/Fortran/Java functions) as subroutines to an IML program, which may simplify greatly some programming task.

Moreover, the great power of its data management instructions (ideally designed for panel data) justifies the use of SAS by itself. With SAS, the most difficult operations on panel data sets like merging, matching... become virtually instantaneous with very little programming. In addition the ODS (Output Delivery Service) permits to save all or parts of the results in files with HTML, RTF (Word), PostScript or PDF formats. These reasons explain why SAS is frequently used over the world, despite its high costs and its basic limitations.

The documentation is huge (for instance, only for SAS/STAT module, 3 volumes, near 4000 pages!) with a terminology and examples coming very often from biostatistics, agronomy, psychology, sociology... Nevertheless, the Web site proposes interesting technical supports with on line documentations, many data files and SAS example's programs. The SAS Users Group International (SUGI) and the SAS discussion list are exceptionnally active and useful.

26.2.5 Stata (v. 9)

Even if Stata is mainly oriented towards econometrics, social science and biostatistics on longitudinal data, version 9[17] has developed significantly on different topics, particularly for panel data estimation. From a general view point, Stata presents three main characteristics:

- First, it is available in many platforms (Windows, MacOS, UNIX...) and in three different versions: Small Stata managing a limited number of observations and variables, Intercooled Stata which is less limited because of the use of virtual memory and Stata/SE adequate for analyzing larger datasets. The Intercooled version that we use in this review has only a maximum matrix size of (800×800) and the number of variables cannot go beyond 2047 in a dataset.
- Second, its interface is simple but perfectible (the results window could be ameliorated); Stata works through a command mode (in the same way in different platforms) and a Graphical User Interface. When working in a command mode,

[16] Exactly as `Data` and `Proc` steps cannot interchange informations with an other `Data` or `Proc` step without the use of a table (or a SAS macro variable).

[17] Stata 10 is now available.

note than Stata's commands are typed in lowercase letters (cf. below for some exceptions); the same apply for variable names but not for file name. This is not really a problem because Stata has an excellent and really unified syntax which can be learned very quickly (except, may be, when using the macro language). In many cases, it is better to include all commands in a (ASCII) do-file; it's the solution we advocate in this review.

- Third, Stata is a modular software, i.e. many commands are in fact contained in (ASCII) ado-files (automatic do-files), written with a powerful (but sometimes unusual) programming language which makes an extensive use of macros. This is the case for many new panel data commands. The main advantage is that many users' specific procedures are available from the Web site. Moreover, a (rather simple, only one window is available) internal editor for ado and do files is provided.

From a panel data estimation's method viewpoint, Stata is very interesting especially because it comes with useful functions. First, linear panel data estimation methods are numerous and easy to implement. The code (a do file) reproduced in Table 26.15 shows how to estimate the Grunfeld model by OLS, within, FGLS methods.

We note that panel data specific instructions begin with an `xt` word. Moreover, in this example, the panel structure is not declared when reading the datafile but it is realized thanks to the use of `i(ident)` option in the `xt` instruction, which announce either a balanced or unbalanced panel dataset. For Stata, a panel dataset is always organized by individual (but it can be reshaped with a `reshape` instruction or with `xtdata` instruction). Other options (available for most commands) are useful:

- `if condition` for applying a command on a subsample defined by a condition (e.g. `if age < 60`),
- `in range` for applying a command on a specific list of observations (e.g. `in 100/200`),
- `weight varpond` for weighting observations according to values taken by `varpond` variable.
- There is also a special syntax `by varlist: command` for applying a command depending on the value taken by one or several variables (e.g. `by ident: xtsummarize`). This is very useful for panel data.

Table 26.15 Stata code # 1 (grunfeld.do)

```
use c:\epd\data\stata\grunfeld.dta          // reading the dataset
regress gi vf vc                            // ols
xtreg gi vf vc, fe i(ident)                 // within
xtreg gi vf vc, be i(ident)                 // between
xtreg gi vf vc, re theta i(ident)           // fgls
xthaus                                      // hausman test
xttest0                                     // BP LM test
```

There are several useful instructions which describe (`xtdes`), summarize (`xtsum`), tabulate (`xttab`), graph (`xtdata, xtline`) variables in a panel data set. Stata allow several estimation methods for static or dynamic panel data models (IV, GMM, 2SLS...). This can be illustrated through the estimation of a dynamic panel data model's on AB data (*cf.* Table 26.16; reading data file and creation of some variables are not reported).

Several topical points can be raised here. First, the symbol `///` is the continuing line operator. Second, applying AB GMM estimator with `xtabond` instruction (or AH estimator with `xtivreg` or `ivreg2` instructions) is very simple. Moreover, system estimators are available with `xtabond2` and `xtdpdsys` instructions. Third, note that in the context of a dynamic estimation, it is necessary to define explicitly the panel structure by `tsset` instruction (another variant implies to use `iis ident` and `tis year` instructions). Creating variables in first differences may be done by, for example, `generate dk = d.k` and we can use in a similar way `f` (forward), `s` (seasonal) and `d` (difference). We must also underline the fact that we define lagged variables by, e.g., `l(0/2).(dk dys)` which creates lags 0,1 and 2 of the two variables These operators may also be combined. For example, `l2d.(k ys)` gives the same results as `l(0/2).(dk dys)`. With many instruments, you will probably need to adjust the `matsize` parameter. For complicated problemss, Stata offers a new and quite powerful matrix programming language (Mata).

For non linear estimation on panel data, Stata enables to proceed in different ways. First, when maximum likelihood is required, it is possible for users to write their own likelihood function and then to call Stata for maximizing it. One of the most interesting features of this ML module is its ability to maximize the log likelihood summing up the individuals components obtained by time aggregation (for instance, as in a random effects probit). This works on a balanced or an unbalanced panel. Moreover, it is possible, but not necessary, to write the first and second derivatives. But, as the Stata language is somewhat specific (intensive use of macro

Table 26.16 Stata code # 2 (ab.do)

```
// reading data file and new variables creation not reported
set matsize 200                          // used by xtabonb
tsset ident year   // panel data structure necessary for xtabond
// AH-d AB, table 5-e pp. 292
xtivreg n l2.n l(0/1).w l(0/2).(k ys) year3-year9 ///
 (l.n = l3.n), fd
 // AH-l with ivreg2 (gmm) : table 5-f pp. 292
ivreg2 d.n l2d.n l(0/1)d.w l(0/2)d.(k ys) year3-year9 ///
    (l1d.n = l3.n), cluster(ident) gmm
// one-step results (not robust)
xtabond n l(0/1).w l(0/2).(k ys) year4-year9, lags(2)
// one-step results + robust : table 4-a1 pp. 290
xtabond n l(0/1).w l(0/2).(k ys) year4-year9, lags(2) robust
// two-step results : table a2 pp. 292
xtabond n l(0/1).w l(0/2).(k ys) year4-year9, lags(2) twostep
```

Table 26.17 Stata code # 3 (grunfeld_ml.do)

```
clear
program drop _all
program define mlere
 version 8.0
 args todo b lnf
 tempvar theta1 z T S_z2 Sz_2 a
 tempname s_u s_e
 mleval `theta1' = `b', eq(1)
 mleval `s_u' = `b', eq(2) scalar
 mleval `s_e' = `b', eq(3) scalar
 quietly {
  gen double `z' = $ML_y1 - `theta1'
  by i: gen `T' = cond(_n==_N,_N,.)
  by i: gen double `S_z2' = cond(_n==_N,sum(`z'^2),.)
  by i: gen double `Sz_2' = cond(_n==_N,sum(`z')^2,.)
  gen double `a' = `s_u'^2 / (`T'*`s_u'^2 + `s_e'^2)
  mlsum `lnf' = -.5*`T'*ln(2*_pi*`s_e'^2) ///
                -.5*ln(`T'*`s_u'^2/`s_e'^2 +1) ///
                -.5*( (`S_z2'-`a'*`Sz_2')/`s_e'^2 ) ///
  if `T'~= .
        }
end
use c:\epd\data\stata\grunfeld.dta
sort ident year
ml model d0 mlere (eq1: gi=vf vc) (s_u:) (s_e:), ///
   max init(1 1 1 1 1,copy)
ml display
```

variables), this may be difficult to do as shown in Table 26.17 (the code is adapted from Gould, Pitblado and Sribney, 2003).

Second, as in the following example (Random Effects probit model), Stata provides many modules (Ado-files) doing automatically ML estimation on some non linear models on panel data: stochastic frontier models, fixed-effects & random effects logit models, random-effects probit models (with a powerful `quadcheck` instruction which checks the sensibility of estimation results to the selected number of quadrature points), random-effects tobit models... We give now (see Table 26.18) an illustration of using such estimation instructions on KW data.

Other general Stata's features should be taken into consideration (for a more complete review, see Kolenikov (2001), Barrachina and Llopis (2002)). Stata may read and write various file formats: Stata, ASCII (very easily), and dBase, MS Access, Excel files... with the ODBC instruction. The graphics module is very good. Stata offers also several instructions (`append`, `merge`, `joinby`...) to merge a dataset with another one. Many instructions (`recode`, `autocode`, `tabulate`...) permit to create dummy variables very easily. The documentation is excellent (13 volumes, may be the best one of all general purposes packages documentation reviewed in this chapter). There is a special volume which is dedicated to cross-

Table 26.18 Stata code # 4 (discrete_choice.do)

```
clear
set memory 10m
version 8
use c:\epd\data\stata\kw.dta              // reading the dataset
// creating the endogeneous var. (0/1)
generate status = (choice==1)*1 + (choice==2)*2 + (choice>=3)*3
generate choice1 = (status==1)*0 + (status==2)*0 + (status==3)*1
// conditional fixed-effects logistic regression
clogit choice1 educ exper expersq black, group(ident)
// random effects logit model
xtlogit choice1 educ exper expersq black, i(ident) re
// random effects probit model
xtprobit choice1 educ exper expersq black, i(ident) re
```

sectional time series models estimation (Stata, 2003), containing very useful technical appendices. The Stata Web site is equally very convincing: it is possible to call directly a dataset by an `http` instruction in a Stata program and the user may find several ado files which can be downloaded, installed and used with Stata. Automatic updating is also available.

26.2.6 TSP (v. 5)

This latest version of TSP (Time Series Processor) has several new features which introduce substantial improvements compared to the later version (4.2) reviewed in the second edition of this volume, some of them being yet available with the version 4.5.

First, two new interfaces are today available: (a) "TSP through the Looking Glass" which is a two windows interface (a program window and an output/log window), (b) "TSP through GiveWin" (GiveWin is an interactive menu-driven graphics-oriented program in which TSP, as some other software Ox, PC-Give…, may be integrated) but with a moderate extra-cost. In this review, we use "TSP through the Looking Glass".

Second, many limitations of the previous versions have been removed: two dimensional matrices are now available, the graphic module has been enhanced, (with GiveWin, it becomes really nice; note too that special graphics for panel data are available in TSP), and the size of the dataset may be extremely large. Nevertheless, the programming language has some limitations (e.g. string management features).

From the point of view of panel data management, a major change is that the `Freq` instruction can manage panel data structure (balanced or unbalanced) and allow to handle missing values, leads and lags. Suppose you have a balanced panel dataset (Grunfeld's data with 10 individuals observed 20 years, 1935–1954), you may write, at the beginning of your program, `Freq(Panel,T=20)`.

Table 26.19 TSP code #1 (grunfeld.tsp)

```
Freq(Panel,Id=ident) ;
? reading an xls file
Read(File='c:\epd\data\excel\grunfeld.xls') ;
Freq(Panel,Id=ident) A; ? A stands for annual data
? Panel data estimation (ols, between, within, fgls)
Panel(Robust,Hcomega=Block) gi c vf vc ;
```

For unbalanced data (for instance, AB data, 140 individuals, 1976–1984), it is necessary to have in the data file one series having a numerical identifier for each individual (for instance, `ident` $= 1,2,3,\ldots,140$). Then, you can define `Freq(Panel,id=ident) A;`.

Linear panel data estimation is mainly done with the usual `Panel` instruction, which is illustrated on the following example using Grunfeld's data. Note that the syntax of Table 26.19 is clear and concise.

The output is very detailed; an F test of fixed effects and an Hausman test are automatically reported.

Several recent improvements are of major interest for panel data models. New estimation commands allow to estimate (1) RE (individual and/or time) model by ML, (2) RE or FE model with a AR1 error term by ML, (3) RE or FE probit model and (4) 2SLS and GMM with FE. When estimation methods and tests on panel data models (linear and nonlinear) are not available with pre-programmed instructions; fortunately, on two TSP companion Web sites, it is possible to find dozens of programs mainly for IV, GMM…panel data estimation. Of course, users have to write programs based on a good matrix language and with a powerful maximum likelihood instruction (look at the `Differ` instruction, for instance). We are going to illustrate these points with the same two examples used previously: a dynamic linear model by AH estimator and a random effects model estimated by ML.

For dynamic linear model on panel data, applying the AH estimator, is easy (in Table 26.20, we do not report code for reading the data file).

Table 26.20 TSP code #2 (ab.tsp)

```
? creating new variables
dn = n - n(-1) ;dw = w - w(-1) ; ? gener is optional
dk = k - k(-1) ;dys = ys - ys(-1) ;
Dummy year ; ? creates year dummies year1...year9
? Panel data estimation : AHd, cf. A&B pp. 292
List lvar dn(-3) dn(-2) dw dw(-1) dk dk(-1) dk(-2)
          dys dys(-1) dys(-2) year5-year9 ;
Select .Not.Miss(dn(-3)) ;
? AHd robust
2sls(Robust,Hcomega=Block,Inst=lvar) dn c dn(-1) dn(-2)
     dw dw(-1) dk dk(-1) dk(-2) dys dys(-1) dys(-2) year6-year9 ;
```

Table 26.21 TSP code #3 (discrete_choice.tsp)

```
Title 'Random effects probit' ;
Probit(Rei) choice C educ exper expersq black ;
Title 'Fixed effects logit' ;
Probit(Fei) choice C educ exper expersq black ;
```

Note the use of the `Dummy` instruction which creates very easily time dummies, denoted `year1...year9`. With the last release, when using the `Robust` and `Hcomega=Block` TSP's options, the SEs are robust to general cross-section heteroscedasticity and time series autocorrelation. Nevertheless, for GMM estimation (AB estimator), there is not a `Hcomega` option, so so we cannot reproduce exactly AB results. On TSP companion Web site, a sample program is available to do this task, but the solution is too complicated (in creating the instruments matrix and computing the var-cov matrix) to be reproduced here.

TSP offers also some instructions for estimating logit/probit models on panel data. For instance, the example of Table 26.21 shows how to estimate random effects probit and fixed effects probit models on KW data.

Note that TSP uses analytical first and second derivatives in these maximum likelihood estimation, which implies that this code is executed very quickly. For other maximum likelihood estimations, there are two ways to specify the log likelihood. First, we can use the `Frml` and `Eqsub` instructions. The major advantage of this method is that TSP can compute first and second analytical derivatives (see above). This method contributes largely to the speed and the possibility to achieve convergence. When it is difficult to write the log likelihood with one or more `Frml` instructions (e.g. with panel data, it is necessary to sum up the log likelihood on time periods), the user may write it with the `Proc` method. If this case, derivatives are evaluated numerically, sometimes not very adequately, and are very time consuming. We illustrate here only the first possibility on the estimation of a linear random effects models on Grunfeld's data (see Table 26.22, some parts of the program are not reported).

Table 26.22 TSP code #4 (grunfeld.tsp)

```
? log likelihood as in Nerlove (2002)
Param beta0,1 beta1,1 beta2,1 sigma2,1 rho,0.5 ;
Frml rei logl = - Log(sigma2)/2 - (0.5/(sigma2))*((e)^2)
          - Log(ksi)/(2*t) - Log(eta)*(t-1)/(2*t) ;
Frml ksi 1 + rho*t - rho ;
Frml eta 1 - rho ;
Frml e   gistar - beta1*vfstar - beta2*vcstar- beta0/sqrt(ksi) ;
Esqub rei e ksi eta ;
Title 'ML Random individual effects' ;
Ml(Maxit=100,Hiter=N,Hcov=N) rei ;
```

More generally, TSP can run on various type of hardware (Windows PC, Apple Macintosh, Unix workstations and some mainframes). The `Sort` instruction admits only one key, and merging panel datasets does not seem not easy. Lastly, cross-tabulation instructions are not available. It can read and write numerous file formats (ASCII, XLS, DBF, Stata...) and also provide some databank management capabilities. For more details, one can refer to Silk (1997), Lubrano (1998), Merino (1998).

The documentation is good, but contains too few examples on panel data, which is not really a problem due to the great amount of explanations available on the Web site. More information may be obtained by contacting the efficient technical support.

26.3 High-Level Matrix Programming Languages

Econometricians use more and more matrix programming language in at least two situations:

1. They need to apply new econometric techniques, not yet available in general econometric packages, and difficult to implement with them,
2. The econometric method is too time-intensive and so, requires a very efficient programming language (Monte-Carlo simulation, estimation by simulation...) without requiring knowledge of a low level programming language[18] (e.g. C or C++).

Matrix programming languages are very often an efficient solution to these difficulties, easier than using low level programming languages. Two packages seem to dominate, at least when panel data estimation is concerned: GAUSS and Ox.

26.3.1 GAUSS (v. 5)

GAUSS[19] is an econometrics and statistics oriented programming language designed to perform statistical and mathematical (specially matrix) calculus. GAUSS is available in two versions: GAUSS for UNIX workstations and GAUSS for Windows. It consists mainly of three components: (1) The GAUSS programming language, (2) a library of application modules (Time Series, Maxlik, GAUSSplot...) and (3) some add-on programs (GaussX...), these last two components available with an extra cost. Due to its speed, to its intuitive and powerful syntax, and above

[18] See Cribari-Neto (1999) and Eddelbüttel (1996) for the interest to use low level programming languages for econometric purposes.

[19] Version 8 is available and offers new or updated libraries (Algorithmic Derivatives, Discrete Choice, Symbolic Tools, GAUSSplot...). There is also a new feature, directly related to panel data models, which allows to use 3 (or more) dimensional arrays. This may be quite useful to manage balanced panel data in arrays with individual, time and regressor dimensions.

all, to the free availability of numerous user's programs (DPD, Expend...), GAUSS is probably the most commonly used of the high-level matrix programming languages, particularly in panel data econometrics (cf. Cribari-Neto (1997), Heywood and Rebelo (1996) and Vinod (2000) for a more general presentation).

For panel data estimation, various strategies may be followed.

For static linear models on panel data, we can first use the TSCS command (included in the Time Series application module - with an extra cost) which may estimate a panel data model by OLS, within and FGLS estimators, providing also an F test of no individual fixed effects and a Hausman test (fixed effects model *vs* random effects model). In Table 26.23, there is an example of TSCS's use on Grunfeld's data.

TSCS it not very sophisticated but is fast and can manage a panel data set (balanced or unbalanced) of whatever size thanks to its working by block of individuals. As TSCS command is written in GAUSS (put in a procedure), then users can ameliorate it by adding its own code for doing new tests, etc.

Another way to estimate panel data models consists in writing on your own a GAUSS program. In this case, the program would be quite different depending on the fact than your panel data set is balanced or not and be of a moderate size or not. With a small balanced panel dataset (for instance the Grunfeld's data), programming OLS, between, within and FGLS is easy, as shown in Table 26.24, where we reproduce a simplified but working example (we limit us to within estimation).

This brief example illustrates GAUSS main characteristics: GAUSS works mainly on matrices and vectors (`xdat`, `x` and `y`). Once, the matrices defined (by extracting columns of `xdat`), computation of an estimator, e.g. OLS, is made by `bols = Invpd(x'x)*x'y;`. For within estimation, it is a little bit more complicated, we must first create $y_{it} - y_{i.}$ and $x_{it} - x_{i.}$. To achieve this result, we must reconfigurate the `y` matrix (and also the `x` matrix) by `ynt = Reshape(y,nn,tt) ;` (`ynt` has now `nn` lines and `tt` columns). Then, individual means are created by `yb = Meanc(ynt') ;` (note the need for transposing because `Meanc` computes means by columns) and finally `yw = Vecr(ynt-yb) ;` computes deviation to individual means to stack in an (nn,1) vector. Lastly, within estimation is done by applying OLS on transformed data by `bwith = Invpd(xw'xw)*xw'yw ;`. This may be completed by computing residuals, tests... In fact, the most tedious part is often to obtain a correct presentation of the output (the `Print` and `Printfm` instructions

Table 26.23 Gauss code #1 (grunfeld_tscs.prg)

```
Cls ;New ; Closeall ;
Library Tscs ; #Include tscs.ext ; Tscsset ;
lhs = { gi } ; exog = { vf, vc } ;
grp = { ident } ; _tsmeth = 1 ;
filename = "c:\\epd\\data\\gauss\\grunfeld";
{bwtih,vbwith,mdv,bec,vbec,mec} = Tscs(filename,lhs,exog,grp) ;
```

Table 26.24 Gauss code #2 (grunfeld_gauss.prg)

```
/* initializations */
Cls ; New ; nn = 10 ; tt = 20 ; nt = nn*tt ;
file = "c:\\epd\\data\\gauss\\grunfeld" ;
/* reading the Gauss data file */
Open f1 = ^file For Read ; vnoms = Getname(file) ;
xdat = Readr(f1,nt) ; f1 = Close(f1) ;
/* xdat is a matrix (200,# of variables in the file)*/
/* defining y and x matrices from xdat */
Let pexo = vf vc ; Let pendo = gi ;
lvexo = Indcv(pexo,vnoms) ; lvendo = Indcv(pendo,vnoms) ;
x = Submat(xdat,0,lvexo)~Ones(nt,1) ; y = Submat(xdat,0,lvendo) ;
/* Within transformation */
ynt = Reshape(y,nn,tt) ; w = Vecr(ynt-Meanc(ynt')) ;
x1nt = Reshape(x[.,1],nn,tt) ; x2nt = Reshape(x[.,2],nn,tt) ;
xw = Vecr(x1nt-Meanc(x1nt'))~Vecr(x2nt-Meanc(x2nt')) ;
/* within */
bwith = Invpd(xw'xw)*xw'yw ; ? bwith ;    /* to be completed */
end ;
```

are complex and not very powerful; this is also true for printing instructions provided by Ox and SAS/IML).

With a large unbalanced panel dataset, the program must be adapted in the following way. For memory space limitations, computations must be done individual by individual. So, it is useful to read first an auxiliary file (say `auxti`) giving for each individual (represented by an identification variable, say `ident`), the number of times this individual is observed, say, a variable named `tii` $= 3,5,8,3\ldots$ for instance. Second, we read the data file[20] (called `mydata`) and do the computations individual by individual (or with more programming by block of individuals). The following code illustrates this idea (note than DPD-GAUSS, but not TSCS, use this principle). We limit ourselves to within estimation due to space limitation. This code is reproduced in Table 26.25.

A `reshape` operation is then not necessary, because we work individual by individual. The program will be slower due to this working (but it may be speeded up if working by block of individuals which implies a complication in programming), and because probably several do loops on the data file will be necessary, one for computing OLS, between and within estimation and at least another one for FGLS estimation, tests, etc.

For dynamic linear model estimation on panel data, fortunately things are simpler because of the DPD-GAUSS program (cf. Arellano and Bond, 1998) which computes estimates for dynamic models from balanced or unbalanced panel data. This program provides several estimators (OLS, within, GMM, instrumental variables,

[20] It is also possible to read all the variables into memory (GAUSS put them in a workspace) and then to do the computations by individual. This will speed up the execution and, at the same time, will save memory.

Table 26.25 Gauss code #3 (simple_by_indiv.prg)

```
New ;
/* reading the auxiliary datafile */
file = "c:\\epd\\data\\gauss\\auxti" ;
Open f1 = ^file For Read ;
nind = Rowsf(f1) ; z = Readr(f1,nind) ;
f1 = Close(f1) ; tii = z[.,2] ;
/* defining variable names */
file = "c:\\epd\\data\\gauss\\mydata" ;
vnoms = Getname(file) ;
Let pexo = x1 x2 x3 x4 x5 x6 x7 x8 x9 ; Let pendo = y ;
lvexo = Indcv(pexo,vnoms) ; lvendo = Indcv(pendo,vnoms) ;
/* read and compute by individual */
Open f2 = ^file for Read ;
i = 1 ; xwtxw = 0 ; xwtyw = 0 ;
Do While i <= nind ;
   data = Readr(f2,tii[i]) ;
   y = Submat(data,0,lvendo) ; x = Submat(data,0,lvexo) ;
  /* between and within transformation */
  xm = Meanc(x) ; ym = Meanc(y) ; xw = x - xm' ; yw = y - ym' ;
  xwtxw = xwtxw + xw'xw ; xwtyw = xwtyw + xw'yw ;
  i = i + 1 ;              /* next individual */
Endo ;
f2 = Close(f2) ;
bw = Invpd(xwtxw)*xwtyw ; ? bw ;
End ;
```

system estimators…) and robust test statistics (Sargan test, tests for serial correlation…). Reading the data by block, DPD is not limited by a maximum number of observations. The instrumental variables matrix can contain several hundred of columns, the main limitation being its invertibility. In order to work with the DPD program, the user has to supply two GAUSS data file: one containing the NT observations of the k variables (sorted by individual and consecutive), and the second one (an auxiliary file) indicating the structure of the main data set. For the AB data,[21] this auxiliary file looks as in Table 26.26.

Table 26.26 Contents of the auxiliary file (abaux.dat)

nbyear	count
7	103
8	23
9	14

[21] These files are provided with DPD-GAUSS. We only change their names.

These two files[22] (named ab.dat and abaux.dat) permit to DPD to read quickly the panel but require that the data file must be sorted by nbyear and by individual. Note that the main dataset should contain a time stratification variable and, optionally, an individual stratification variable. The next step is to modify, with the GAUSS editor, an (ASCII) file: DPD.RUN (you can change its name). This file contains GAUSS instructions which define the name and path of main and auxiliary datasets, the name of time, individual stratification, independent and dependent variables and the definition of instruments... This program calls for two related programs (located in the Gauss directory): DPD.FNS containing some functions for data transformations and DPD.PRG, the main program file. DPD.RUN may look as in Table 26.27 (instructions used only for printing, unmodified lines and some comments are not reproduced).

This code allows to reproduce AB results (see Arellano and Bond (1991), Table 4-a1, a2). One of the crucial points is the choice of the instrumental variables and the DPD function named GMM() which returns the optimal instrument matrix for the GMM estimators. This matrix may be combined with other matrices using the vertical concatenation operator (~). DPD-GAUSS allows also to use system GMM estimators combining moment conditions for equations in first differences with moment conditions for equations in levels (with the functions Lev1(), Diflev()... Another major interest of DPD-GAUSS is that the authors provide the source code, so you can adapt it to your own problem if required. If you want to estimate your model by OLS or within, you need only to adapt the

Table 26.27 Gauss code #5 (ab_gauss.prg)

```
bat=1;                                  /* mode batch */
imod=1;                                 /* model in diff. */
icon = 1 ; irob = 1 ;  /* constant and robust estimation
choices */
open f1="c:\\epd\\data\\gauss\\xdata";   /* main data */
open f2="c:\\epd\\data\\gauss\\auxdata";       /* auxiliary data */
yearcol=2;                  /* Data column for year */
year1=1976;                    /* First year of data */
nyears=9;               /* Number of years in data set */
lag=2;            /* Longest lag to be constructed due to n_{t-2}*/
                   /* 2 first obs. by indiv. skipped */
data=ln(data);               /* all variables in log */
y=dif(3,0);                /* endogeneous var. */
/* exogeneous var: note the use of the diff function */
x=dif(3,1)~dif(3,2)~dif(4,0)~dif(4,1)~dif(5,0)~dif(5,1)
~dif(5,2)~dif(6,0)~dif(6,1)~dif(6,2);
/* instruments definition */
z=gmm(3,2,99)~x[.,3:10];   /* note the use of gmm function */
/* more lines follow */
```

[22] Recall than with GAUSS, a dataset is made, in general, of two files: one with the extension .dat containing the data, the second one with the extension .dht, giving the columns' names. So, in our example, four files are used: ab.dat, ab.dht, abaux.dat and abaux.dht.

choice of the instruments matrix in writing `z=Ols ;` (or `z=x ;`) for OLS and `z = Wgroups ;` for within estimator. To be used correctly DPD requires a bit of knowledge of GAUSS and a little idea of DPD working (individual by individual, all the variables are in the matrix called here `data`). A careful reading of the user's guide for DPD (Arellano and Bond, 1998) is absolutely necessary.

For non linear or maximum likelihood estimation, three application modules are very useful: constrained optimization (CO), maximum likelihood (Maxlik) and constrained maximum likelihood (CML) modules. We are going to illustrate the use of Maxlik on a simple example,[23] estimation of an error components model by ML on Grunfeld's data (data reading and variables creation not reproduced), see Table 26.28.

It is clear according to this example that GAUSS programming supposes a good knowledge of basic programming concept (procedures, local and global variables...).

The GAUSS documentation is good, but contains too few examples. The Aptech web site is mainly limited to commercial information. Fortunately, there are numerous users programs which can be easily founded and used, e.g. for panel data estimation:

Table 26.28 Gauss code #6 (grunfeld_ml_gauss.prg)

```
New ; Closeall ; Library Maxlik ; #Include Maxlik.ext;

/* ML estimation of error components model */
xb = xb.*.Ones(tt,1) ; yb = yb.*.Ones(tt,1) ;
maxl = yb~Ones(nt,1)~xb~yw~xw ;
Maxset ; Maxclr ;
x0 = 1~1~1~1~0.5 ; /* initialization values */
{bml,logl,g,h,retc} = Maxlik(maxl,0,&logl,x0) ;

Proc logl(b,z) ;
Local llog , ksi , eta , part3 ;
ksi = 1 + b[5]*tt - b[5] ;
eta = 1 - b[5] ;
part3 = ( z[.,5]/Sqrt(eta) + z[.,1]/Sqrt(ksi) )
  - b[2]*( z[.,3]/Sqrt(ksi) + z[.,6]/Sqrt(eta) )
  - b[3]*( z[.,4]/Sqrt(ksi) + z[.,7]/Sqrt(eta) ) -
b[1]/Sqrt(ksi) ;
llog = -nt*Ln(2*PI)/2 -nt*Ln(b[4])/2 -nn*Ln(ksi)/2 -
(nt-nn)*Ln(eta)/2
  -(1/(2*b[4]))*(part3'*part3) ;
Retp(llog) ;
Endp ;
End ;
```

[23] A more complex one (random effect probit model estimation) is given in `reprobit_gauss.prg`).

- C. Kao and I. Choi provide free programs for unit root and cointegration tests for panel data,
- Many GAUSS programs on linear and non linear (binary, censored, count, sample-selection...) models are provided by Lee (2002),
- Train (2003) provides free code for mixed logit estimation for panel data,
- Windmeijer (2000) offers a GAUSS program for non-linear GMM estimation of exponential models with endogenous regressors for cross section and panel (dynamic) count data models (see also Romeu, 2004).

26.3.2 Ox (v. 3.4)

Ox is a true object matrix programming language available for many platforms (Windows, DOS, Unix, Linux,...).

All Ox versions are free for academic use, except the Windows version. The free versions are named OxConsole (for DOS/UNIX) as they are called by a command line in a console windows. An editor is not provided with this basic version which cannot moreover visualize graphs (but it can save them in a Postscript file).

The Windows version (called Ox Professional[24]) may use two types of interface: first, GiveWin, and, secondly, OxEdit. OxEdit (provided with a purchase of Ox) is a text editor developed, as for Ox, by J. Doorniks. In this review, we use Ox Professional. (v. 3.3) for Windows with OxEdit. OxEdit has two interesting features: first, when OxEdit edits an Ox source code, it uses colors[25] to distinguish between instruction keywords, comments, numbers, syntax errors...; second, it may be used as a front-end to Ox (but also to C, C++, TEX, LATEX...). Nevertheless, an advantage of GiveWin over OxEdit is that it allows users to modify a graph created by Ox by adding text, labels,... Moreover, GiveWin enables to manage data (editing, variable creation...) more easily.

The Ox language has several specific features:

- It has broadly a similar syntax to C, C++ and Java (e.g. Kerninghan and Ritchie, 1998, Stroustrup, 1992). All indexing of matrices and vectors start at 0 and not at 1. The main difference from C, C++ is that a matrix is a standard type in Ox. So, even if all variables must be declared before use, their type is defined only implicitly. Therefore, a variable may start as an integer and then be redefined as a matrix.
- As GAUSS, it has an extended graphical, mathematical and statistical functions library and similar matrix operator (concatenation, inversion...). Ox also allows

[24] Ox 4.1 is now available. Note that we do not review another well-known econometric software, PcGive for two reasons: (1) space limitations and, (2) PcGive is mainly written in Ox language and shares several features with Ox. Nevertheless, it uses pull-down menus and is a simple and efficient alternative to Ox.

[25] This is also true for GiveWin and for the SAS Enhanced Editor.

for vectorized code, more efficient for programming and execution, and very often, Ox and GAUSS syntaxes are similar. Nevertheless, in some cases (Do-loops...), Ox syntax may be unusual for non C, C++ users.
- Nevertheless, Ox syntax is case sensitive and many instructions must be written in lowercase, some exceptions arise when using Ox classes (see below). Variable names cannot exceed 60 characters (the first one being a letter).
- Ox can read various datafile formats: PcGive/GiveWin, XLS, GAUSS, Stata, ASCII...

Some of these aspects may be illustrated through a short example (closed to the GAUSS[26] one, *cf.* pp. xxix), given in Table 26.29.

In fact, the main difference between GAUSS and Ox is the concept of Class. Ox provides pre-programmed classes,[27] such as the Database class (used to store data with database functionalities). One of the major interest of the class concept is that it avoids using global variables. It is probably with this concept of classes (an optionally feature) that a non C,C++ user could encounter trouble. Using classes is very useful to develop professional applications. Moreover, if creating Ox class may be difficult,[28] using it is not, as we can see in Table 26.30 which illustrates some uses of the Database class (used by many other classes).

Table 26.29 Ox code #1 (grunfeld1.ox)

```
#include <oxstd.h>
main()
{
 decl data, x, y, nt, nn, tt, ynt, yb, yw,
 x1nt, xb1, x1w, x2nt, xb2, x2w, xw, bwith ;
 data = loadmat("c:\\epd\\data\\excel\\grunfeld.xls");
 nt = 200 ; nn= 10 ; tt = 20 ;      // balanced panel data set
 y = data[][2] ; x = data[][3:4] ;   // defining y and x matrices
 ynt = reshape(y,nn,tt) ; yb = meanr(ynt) ; yw
 = vecr(ynt-yb) ;
 x1nt = reshape(x[][0],nn,tt) ; xb1 = meanr(x1nt) ; x1w
 = vecr(x1nt-xb1) ;
 x2nt = reshape(x[][1],nn,tt) ; xb2 = meanr(x2nt) ; x2w
 = vecr(x2nt-xb2) ;
 xw = x1w~x2w ; bwith = invertsym(xw'xw)*xw'yw ;
 println("bw = " , bwith) ;
}
```

[26] Note also that it is possible to run a GAUSS program under Ox, see Doornik (2004), Laurent and Urbain (2004) and Viton (2003). Of course, Ox can be interfaced with C, C++ programs.

[27] Following Podovinsky (1999), a class provides "an abstract definition of an object (both in terms of variables - the data - and functions - the methods that apply to the data-)".

[28] A user interested in writing his own Ox classes may refer to Doornik, Draisma and Ooms, (2001,Chap. 8, pp. 78–91).

Table 26.30 Ox code #2 (grunfeld2.ox)

```
#include <oxstd.h>  #import <database>   // don't put a ;
main()
{
decl mydb, mnames, names1, y, ly, mdata, ident, year ;
mydb = new Database();
mydb.Load("c:\\epd\\data\ox\\grunfeld.in7");   // load data
mydb->Info() ;                    // info on the file
mnames = mydb->GetAllNames() ;   // load var. names into a vector
println("# of obs = " , mydb->GetSize() ) ;   // print nb. of obs.
println("var. names = " , mnames ) ;    // and var. names
mdata = mydb->GetAll() ;    // load all data in a matrix
ident = mydb->GetVar("ident") ;    // load ident var. into ident
mydb->SetSelSample(-1,1,-1,1) ;    // set sample, necessary

year = mydb->GetVar("year") ;    // load year var. into year
mydb->Select(0, {"gi",0,0, "vf",0,0} ) ;   // 2 var. in group 0
println("# of indiv : ", columns(unique(ident)),
  "\n period : ", min(unique(year)) ,
    " : ", max(unique(year)) ) ;
y = mydb->GetGroup(0) ;    // load 2 var from group 0 into y
ly = log(y) ;   // taking log of 2 variables
names1 = {"y", "log(y)"} ;       // print result for checking
println("%c", names1, (y~ly)[0:1][]) ;  // 2 rows, all columns
delete mydb;          // finished with object
}
```

This short program suggests two points:

- With the Database class, it is easy to read (and to create) a data file with `mydb->Load(...)`; Moreover, `mydb->Info()` ;may be used to obtain some file information (variable names, number of missing values...).
- We see according to this example that this class enables to access to the data but also to predetermined functions working on this data (which may be different from the class used). For instance, the `GetSize()` function returns the number of observations in the file. `GetVar(...)` and `GetAll()` allow to put some or all variables in a vector or in a matrix.
- In the same way, with `Select(...)` and `GetGroup(...)`, we can define some object (a group named 0, 1 containing one or several variables...) and then manipulating it.
- Don't forget to define the `SetSelSample(...)` which is mandatory. Remember that missing values are dropped out. Note also, that `SetSelSample(...)` must be defined just after the `Select(...)` instruction, but only if you use `Select(...)` and `GetGroup(...)...`
- There are many functions which may be used for each class. Note however that the Database class is not, as it stands here, able to manage panel data if, for

instance, we want to create lags (and to put a missing value for the first observation for each individual). This can be realized by an other class, DPD class which has many of the Database class instructions (plus some other ones, that we will examine later) or by hand (in a loop).

Ox comes with several (free) packages or classes:[29] Maximization package (numerical optimization and differentiation), Probability package (density, cumulative density, random number...), Simulation class (Monte Carlo experiments), QuadPack (numerical integration), Lapack (matrix calculus)... Among these packages, three may be useful for panel data estimation.

Ox-DPD (cf. Doornik, Arellano and Bond, 2002) is a package for estimation of dynamic panel data models, whose functionalities are closed to DPD GAUSS. In Table 26.31, we illustrate Ox-DPD use by an example.

Table 26.31 Ox code #3 (ab.ox)

```
#include <oxstd.h> #import <packages/dpd/dpd>
main()
{
decl dpd = new DPD();   // defining data set and model used
dpd.Load("c:\\epd\\data\ox\\abdata.in7");   // load data
dpd.SetOptions(TRUE); // robust standard errors
dpd.SetYear("YEAR"); dpd.SetGroup("IND");
dpd.SetDummies(D_CONSTANT + D_TIME);    // specify dummies
dpd.Select(Y_VAR, {"n", 0, 0});    // endogeneous & regressors
dpd.Select(X_VAR, {"n", 1, 2, "w", 0, 1, "k", 0, 2, "ys", 0, 2});
print("\n\n***** Within *****");
   dpd.SetTransform(T_WITHIN); dpd.Estimate();
print("\n\n***** AH diff *****");
   dpd.Select(I_VAR, {"n", 2, 3, "w", 0, 1, "k", 0, 2,
"ys", 0, 2});
   dpd.SetTransform(T_DIFFERENCES); dpd.Estimate();
print("\n\n***** AB col a2 *****");
   dpd.DeSelect() ;  // reformulate model, warning, see the text
   dpd.SetYear("YEAR"); dpd.SetGroup("IND");
   dpd.SetDummies(D_CONSTANT + D_TIME) ;
   dpd.Select(Y_VAR, {"n", 0, 0});
   dpd.Select(X_VAR, {"n", 1, 2, "w", 0, 1, "k", 0, 2,
 "ys", 0, 2});
   dpd.Select(I_VAR, {"w", 0, 1, "k", 0, 2, "ys", 0, 2});
   dpd.Gmm("n", 2, 99);   // GMM-type instrument
   dpd.SetTest(1, 2); // Sargan,AR 1-2 tests
   dpd.SetMethod(M_2STEP); dpd.Estimate();   // 2-step estimation
delete dpd;   // finished with object
}
```

[29] Loosely speaking, there is a difference between a package and a class: a package is a compiled version of one or more classes.

We can make two comments about this code.

- The panel structure is declared with the `dpd.SetYear("YEAR")` instruction. So you do not need an individual identifier in your datafile (Ox-DPD creates an index variable by differencing the year variable). If you have one, it is better to write also `dpd.SetIdent("IDENT")` when, e. g., for some individuals the last year is 1998 and the first one of the next individual is 1999. One tricky point is that you must not confuse uppercase (`''IDENT''`) and lowercase (`''ident''`) depending upon the way you create the variable.
- Be careful with the statements `dpd.Select(...), dpd.Gmm(...)`... For instance, if you write `dpd.Select(X_VAR, {"n", 0, 1});` and after `dpd.Select(X_VAR, {"k", 1, 2});` there will be 4 variables in the regressors list, $\mathtt{n}_t$, $\mathtt{n}_{t-1}$, $\mathtt{k}_{t-1}$ and $\mathtt{k}_{t-2}$. This explains why you will need the `dpd.DeSelect();` which clears completely the list.

Second, with the the Maximization package, it is easy (and in a very fast manner) to estimate a random effects model by ML on Grunfeld data, as shown in Table 26.32. Note two points: 1°) GAUSS offers broadly the same functionalities, but Ox does not provide a constrained optimization and a constrained maximum likelihood modules; 2°) If the log likelihood has to be evaluated by individuals, the program becomes more more complicated (see `probitreml.ox` for an illustration).

Lastly, we must mention also a new (free) package, DCM (Discrete Choice Models) written by Weeks and Eklöf (see Eklöf and Weeks 2004a, b). DCM is a package to estimate several classes of discrete choice multinomial models, particularly, conditional logit, mixed logit, multinomial probit, nested logit, ordered probit... on cross-section and (in somes cases) on panel data. DCM is a class written in Ox and its use is simple and flexible. This is especially true concerning[30] the structure of the database to be read (which consists in observations about individual and/or choice characteristics for a combination of individuals, time and alternatives). DCM can read several structures with few and simple instructions. As, up to now, in the version 1.0, all panel data features are not yet implemented (only for conditional logit and mixed logit model), we do not give more details on it (see `discrete_choice.ox` for a simple example), but it is a promising package.

As a matrix programming language, Ox competes directly with GAUSS for several reasons: First, Ox is cheaper than GAUSS (recall that some Ox versions are free); Second, if you are a C or C++ users, the object-oriented programming approach gives more flexibility; Lastly, in many cases, Ox is faster than GAUSS (see Steinhaus, 2002, Küsters and Steffen, 1996 and Sect. 26.4.1 for an evaluation of speed in a panel data context). For more general reviews, cf. Cribari-Neto (1997), Cribari-Neto and Zarkos (2003), Kenc and Orszag (1997) and Podovinsky (1999).

Ox has an excellent (but quite technical) documentation (see Doornik, Draisma and Ooms (2001) and Doornik (2001)). The Web site is excellent (Doornik's one, because Timberlake web site is only commercial) with the possibility to consult or

[30] It is also easy to store the estimation results in a LaTeXformatted table.

Table 26.32 Ox code #4 (chronoml.ox)

```
#include <oxstd.h> #include <oxfloat.h> #import <maximize>
decl gY, gX, gTT ; // global data
LOGLreml(const coeff, const mloglik, const
 avScore, const amHessian)
{
decl llognc ;
llognc = -log(2*M_PI)/2- (gTT-1)*log(coeff[3]^2)/(2*gTT)
  - log(coeff[3]^2+gTT*coeff[4]^2)/(2*gTT)
  - (1/(2*coeff[3]^2))*((gY[][1]- coeff[1]*gX[][2]
  - coeff[2]*gX[][3]).^2) - (1/(2*coeff[3]^2+gTT*2*coeff[4]^2))
  *((gY[][0]- coeff[0]- coeff[1]*gX[][0]- coeff[2]*gX[][1]).^2 );
mloglik[0] = double( meanc(llognc) ); return 1;
}
main()
{
decl coeff, valfunc, coder, data, x, y, nt, nn, tt, yb,
 yw, xb, xw,
 mhess, mcovar ;
 data = loadmat("c:\\epd\\data\\excel\\grunfeld.xls") ;
 y = data[][2] ; x = data[][3:4] ; // y, x
 yb = data[][5] ; xb = data[][6:7] ; // yb, xb
 yw = data[][8] ; xw = data[][9:10] ; // yw, xw
 tt = 20 ; gY = yb~yw ; gX = xb~xw ; gTT = tt ;
 print("Random effects model by ML","\n\n");
 coeff = <-42.71; 0.115; 0.23; 94; 94>; // ols starting values
 coder = MaxBFGS(LOGLreml, &coeff, &valfunc, 0, TRUE) ;
 Num2Derivative(LOGLreml,coeff,&mhess)   ;
 mcovar = invertgen(-mhess) / rows(y) ;
 print("parameters:", coeff' ,"\
nstd err.:", sqrt(diagonal(mcovar)) ) ;
}
```

download the Ox documentation. The Ox-users discussion group is not very active but the Doornik's support is very efficient.

26.4 Performance Hints and Numerical Accuracy Evaluation

26.4.1 Speed Comparison[31]

First, in Table 26.33 we provide, for each software, time execution[32] to estimate a linear panel data model with 10 regressors by OLS, within and FGLS estimators on a

[31] More detailed speed evaluations (but not on panel data) can be found in Küsters and Steffen (1996), Nerlove (1998) and Steinhaus (2002).

[32] In this review, we use a Pentium 4, 1.6 GHz, with 70 Go HDD and 256 MB RAM running under Microsoft Windows 2000.

Table 26.33 Speed evaluations

Software	Linear Model Data Reading Time[a] +Execution Time	ML Estimation Results[b]
EViews	7s	Marquardt (s2,160,5.2s), BHHH (s2,200,6.6s)[c]
GAUSS	TSCS: 2s User's program: 1.7s	BFGS(s1,124,0.23s), DFP(ns) Newton (s2,77,0.86s), BHHH (s2,967,3.64s)[d]
LIMDEP	4s	BFGS (ns), DFP (s2,16,3.8s) Newton (s1,34,36.38s), BHHH (ns)[e]
Ox Pro.	DPD/Ox: 5s User's program: 1.3s	BFGS (s1,247,0.17s)[f]
RATS	5s	BFGS (s1,117,2s), BHHH (s1,930,35s)[g]
SAS	10s	Newton(s2,5,0.06s), Gauss (s1,207,0.51s) Marquardt (s1,81,0.18s)[h]
Stata	17s	Newton-Raphson (s1,11,1.25s)[i]
B		HHH (s2,937,8.5s), Newton (s1,28, 0.25s)
TSP	7s	Gauss(s1,28,1s), BFGS (s2,32,0.16s) disCrete hessian (s1,28,1.08s)[j]

[a] The files were saved in the proprietary's format for each software. File sizes are about 8-15 Mb.
[b] For each available algorithm, we give between parenthesis (1) the convergence status (s1) (resp. s2) convergence with the first (resp. second) starting values set. Ns indicates no solution, (2) , the number of iterations and (3) the execution time in seconds.
[c] with the LogL object and the ml instruction.
[d] with the Fastmax procedure.
[e] with the Maximize instruction.
[f] with the Maximization package and the MaxBFGS function. [g] with the maximize instruction.
[h] with Proc Nlin.
[i] with the ml command.
[j] with the ml instruction.

large[33] unbalanced panel data set. It contains 10.000 individuals observed between 5 and 12 years (85.160 observations), randomly generated (as in Nerlove, 1971). Second, in the same table, we indicate time performance and convergence status to estimate a random effects model by the maximum likelihood method (see Hsiao, 1986 for the definition of the log-likelihood[34]). Grunfeld's data are used.

[33] In several cases, it is probably necessary to adjust the memory size: for TSP, using OPTIONS instruction, for Limdep using TOOLS/OPTIONS/PROJECT menu, for SAS in modifying the CONFIG file and setting `memory 10m` for Stata.

[34] See Nerlove (2002) for alternative specifications.

Consider, first, the linear model estimation case. In general, speed performances are good and relatively closed. Ox seems to be the faster, and does a little better than GAUSS when using a specific program. Nevertheless, if we use Ox with the DPD class, the program is slower but the results are more detailed, and more estimation methods and options are available. LIMDEP, RATS and TSP are also quite fast when estimating linear panel data models. The slower ones are Stata and SAS. For SAS, it is due to its numerous read/write operations. There is no apparent explanation for Stata. Probably, using the most powerful version of Stata (Stata/SE) would reduce time execution for this software. Note also that GAUSS and Ox are more complex to use in comparison with the other packages. Finally, we can remark that the estimates obtained (not reported here) are very closed between software. The main differences come from FGLS estimation due to different computations of θ_i.

Nevertheless, there are several special cases. With RATS, if we use the `Pregress` instruction, execution time grows to 11mn (due to the estimation of the N fixed effects). It is then better to use the `Panel` instructions (cf. pp. xiii in this chapter) with large panels. For SAS, for the same reason, it was impossible, with my configuration, to apply `Tscsreg` procedure. SAS reported an unsufficient memory error message even when the virtual memory was largely increased. So we use the `Glm` procedure to estimate the FE model and `Proc Mixed` for the RE model. Nevertheless, this last instruction is very inefficient because the estimation is done by ML. Hence, we program directly all the variable transformations and use OLS on transformed variables in order to implement the between, within and FGLS estimators. For GAUSS, we use first the TSCS module and also a specific program we wrote. This program loads all the data in memory and does two Do-loops individual by individual for the computations. So, for all these reasons, in some cases, comparing performance hints is difficult.

General lessons for the maximum likelihood estimation are also difficult to draw. First, note that we try to "standardize" the setup for ML estimation for all software. The likelihood function is the same in each case without specifying first or second derivatives. We do not use any built-in instruction owing to estimate directly RE model by ML without defining explicitely the log-likelihood. As far as it was possible, the convergence criterion is based on the percentage changes in coefficients and/or on the norm of the gradient (both equal to 10^{-5}). Finally, we define two sets of starting values for the coefficients: first very far ones from the solution (labelled as s1); second, s2 is defined according to the OLS estimation. At this point, several comments may be done. First, once again, Ox performs extremely well (quick convergence with s1) but it offers only one algorithm, BFGS.[35] RATS and Stata have also good convergence properties: they converge quickly with s1 (with the BFGS method for RATS). All other software achieve convergence, and sometimes very quickly, sometimes even with s1, but not for all the available algorithms. Nevertheless, EViews converges only with s2. Finally, the most surprising result is that

[35] Newton algorithm may be used but only when providing, at least, first derivatives.

a given method may perform very well with a software but not with another one. So, users have to check their final estimation results in using different algorithms, different starting values and (if possible) different software.

26.4.2 Numerical Accuracy Evaluations

Econometricians have devoted considerable efforts to develop better or new estimation methods. Less was done in controlling numerical accuracy of computational routines used for implementing them. In applied works, panel data users encounter frequently errors or warning messages such as not invertible matrix, not convergence, underflow, overflow... without knowing if these problems come from the data or the algorithm used.

Computational errors arise from a variety of reasons:

- Inexact binary representation of a real value (e.g. 0.1 may be in fact treated as 0.099999964),
- Rounding errors because of limited computer's word length,
- Algorithmic errors (truncation, bugs or inadequate algorithm).

Most of the time (see McCullough and Vinod, 1999,McCullough, 1998,1999a, 1999b), software numerical accuracy is evaluated by comparison with the benchmark results for the Statistical Reference data Sets (SRD) provided by the National Institute of Standards and Technology (NIST, cf. Gill, Guthrie, Lagergren and Rogers, 2001). These reference datasets are mainly artificial datasets created to evaluate software with ill-conditioned data in four general areas useful for our purpose: univariate statistics, analysis of variance, linear regression and non linear regression. Moreover, they are classified according to their difficulty level (lower, average, higher) depending on the number of constant leading digits (3, 7 or 13). NIST provided, for each case, certified values up to 15 (11 for nonlinear estimation) significant digits (*i.e.* the first nonzero digit and all succeeding digits).

We present, for each software, in Table 26.34 a synthesis of NIST benchmark tests. For each test category and for each difficulty level (in the following order: lower, average and higher), we give the mean, and between parenthesis, the minimum and the maximum of number of digits accuracy obtained in using all the datasets for a given level. This number (called LRE) is between 0 (fail) and 15 (perfect result up to 15 digits. It is frequently admitted[36] that, for a lower level of difficulty, LRE must be equal or superior to 10 (6 for nonlinear estimation), whereas for high level, a LRE value equal or greater than 4 or 5 is reasonable.

[36] See, for example, Nerlove 2001.

Three main conclusions may be drawn on these results:

1. First, these results should be taken with care. As there are obtained on very special datasets, their interest for usual econometric practice is controversial. For instance, note that the dataset (SmLs09.dat, high difficulty level) used for checking Anova computations contains 18.009 observations like 1000000000000.n where n equals 2, 3, 4, 5 or 6!. For non linear estimations, the variety of available algorithms and the different settings of convergence criteria make some comparisons quite difficult (see Lilien, 2000). Nevertheless, an important point consists in the fact that all software send a warning or an error message when colinearity or non convergence problems occur.
2. For univariate statistics, Anova and linear regression, it seems that we can be relatively confident in the results obtained by all our reviewed software. More precisely, univariate statistics are very accurately computed. This is also true for standard errors (about 13 digits) but less for the first order autocorrelation coefficient: EViews and RATS fail to compute it correctly in some cases (6 and 3 times respectively). For anova, EViews has a 0 score 3 times. For the most difficult case (smls09), only Ox, SAS, Stata and TSP obtain 2–4 correct digits. Linear estimation results are very closed and good with one exception: with the Filippeli dataset, all software whose results are reported, except LIMDEP and Ox,[37] obtain a zero score.
3. Benchmark results for non linear regression are more mitigated and vary considerably upon the software used and the tests done. Two main comments may de done: First, SAS, Stata and TSP have excellent performances (never less than 5–6 digits, most of the time about 8–10 correct digits) but EViews and RATS works also well even if they fail in some difficult cases. Second, the more surprising result is the poor performance of Gauss and Ox. For some datasets, this may be explained by the fact we do not provide analytical first derivatives. But SAS and Stata also use numerical derivatives. We must note that we use the software defaults for non linear estimation, setting only the maximum number of iterations to 1000 and the convergence criteria to 1e-11. So, at least for Gauss which offers many options (algorithm...), it is possible to obtain better results in modifying the non linear setup. Nevertheless, it is doubtful that the results will be completely changed (see Vinod, 2000).

We conclude this section by giving some information on uniform random generation numbers. In Table 26.35, we summarize the main characteristics of uniform random generator number used by the software that we review. We give the results of a very well known RNG test, called DIEHARD test which consists in 18 specific tests.

We can make several remarks. First, it must be noticed that all analyzed software except RATS and SAS, provide recent RNG. Indeed, RNG with a period of 2^{32} (or less) are useless for large simulation work. For instance, admit that

[37] Another exception is Mathematica, cf. Nerlove, 2001.

Table 26.34 Numerical accuracy results

Software[a]	Univariate Statistics[b]	Anova[c]	Linear Regression[d]	Non Linear Regression[d,e,f]
	15 (15–15)	9.1 (0–15)	12.8 (12.2–13.3)	9.6 (8–11)
EViews 5[g]	15 (15–15)	9.2 (0–15)	15 (15–15)	8 (0–11)
	15 (15–15)	2.2 (0–3.3)	6.4 (0–9.9)	8.5 (0–10.9)
	15 (15–15)	13.4 (12.4–14.5)	11.7 (11.3–12.1)	5.9 (3–7.8)
GAUSS 5[h]	14.5 (14–15)	7.9 (6.5–8.5)	14.9 (14.7–15)	3 (0–6.4)
	14 (14–14)	1.6 (0–2.7)	6.4 (0–13.4)	4.6 (0–7.8)
	15 (15–15)	13.2 (12.4–14)	13.3 (13.2–13.3)	8.9 (7.1–9.9)
LIMDEP 7[i]	14.5 (14–15)	7.7 (6.2–8.4)	14.9 (14.7–15)	7.3 (0–10.6)
	14 (14–14)	1.4 (0–2.4)	10.2 (6.7–14.6)	8.3 (7–9.5)
	15 (15–15)	14.5 (12.9–15.0)	13.1 (12.7–13.6)	5.2 (0–9.8)
Ox Pro.[j]	15 (15–15)	9.9 (9.3–11.7)	14.9 (14.7–15)	3.2 (0–9)
	15 (15–15)	3.3 (3.3–3.3)	9.8 (7.3–12.8)	4.5 (0—-10.5)
	14.5 (14–15)	13.6 (12.4–14.4)	11.8 (11.5–12.1)	6.7 (5.5–8.8)
RATS 6[k]	14 (14–14)	7.9 (6.5–8.5)	14.9 (14.7–15)	5.8 (0–10.6)
	15 (15–15)	1.6 (0–2.7)	6.2 (0–9.1)	6.7 (3.4–9.4)
	15 (15–15)	13.6 (12.7–15)	11.9 (11.5–12.3)	10 (7.4–11)
SAS 9.1[l]	14.5 (14–15)	9.9 (8.8–10.4)	14.9 (14.7–15)	9.8 (6.7–11)
	14 (14–14)	4.3 (4.2–4.4)	6.3 (0–9.6)	9.6 (7.6–11)
	15 (15–15)	13.7 (13.1–15)	12.1 (11.5–12.8)	8.5 (6.7–9.4)
Stata 8.1[m]	15 (15–15)	10.2 (10.2–10.4)	14.9 (14.7–15)	8.1 (4.8–10.9)
	15 (15–15)	4.3 (4.2–4.4)	6.9 (0–12.1)	7.2 (6–8.3)
	15 (15–15)	13.7 (12.3–14.7)	12.5 (12.1–12.9)	9.7 (7.9–11)
TSP 4.5[n]	14.5 (14–15)	10.25 (10.2–10.4)	14.9 (14.7–15)	9.7 (6.5 –11)
	14 (14–14)	3.7 (2.1–4.6)	8.4 (0–12.8)	9.4 (7.9–11)

[a] Author's computations in all cases.
[b] For univariate statistics, we report only the results for mean computation.
[c] For anova, we report only the results for the F result.
[d] For linear and non linear regression, we report only the results for the estimated coefficients.
[e] When non convergence was encountered with the first set of starting values, we provide the results obtained with the second set of initial values.
[f] EViews and TSP use analytical first derivatives. The other softwate use numerical first derivatives.
[g] Used instructions: stats, testby and ls.
[h] Used instructions: meanc, stdc, ols and the Constrained Optimization module. For anova, a specific code is used (see Vinod, 2000).
[i] Used instructions: dstat, xvcm, regress and nlsq
[j] Used instructions: meanc, varc, acf, olsc and MaxBFGS. For anova, a specific code is used (see Vinod, 2000).
[k] Used instructions: statistics, correlate, pstat, linreg and nlls.
[l] Used instructions: proc means, proc arima (+ proc autoreg), proc anova, proc reg and proc nlin.
[m] Used instructions: summarize, corrgram, anova, reg and nl.
[n] Used instructions: msd, bjident, olsq and lsq.

Table 26.35 Random generator number

Software[a]	Function Name	Reference	Max. Period	No. of success to Diehard tests[b]
EViews	`rnd`	Knuth (1997)	2^{129}	18/18
		L'Ecuyer (1999)	2^{319}	18/18
		Matsumoto-Nishimura (1998)	2^{19937}	18/18
GAUSS	`rndu`	Kennedy-Gentle (1980)	$2^{32}-1$	8/18
	`rndKMu`	Marsaglia (2000)	10^{8888}	17/18
LIMDEP	`rnu`	L'Ecuyer (1999)	2^{191}	18/18
Ox	`ranu`	Park-Miller (1988)	2^{32-1}	17/18
		Marsaglia (1999)	2^{60}	17/18
		L'Ecuyer (1997)	2^{113}	18/18
RATS	`%uniform`	Park-Miller (1987)	2^{32-1}	16/18
SAS	`ranuni` `uniform`	Fishman and Moore (1982)	$2^{31}-1$	14/18
Stata	`uniform`	Marsaglia (1994)	2^{126}	18/18
TSP	`random`	L'Ecuyer (1990)	2^{31-1}	17/18
		L'Ecuyer (1999)	2^{319}	18/18

[a] author calculations.
[b] Diehard tests return a p-value. A test fails when the p-value is ¡ 0.01 or ¿ 0.99. When a test returns several p-values, it fails if more than 2 p-values are ¡ 0.01 or ¿ 0.99.

we follow Knuth's recommendation, i.e. set the maximum number of draws as $d < period/1000$. So, in using a RNG with a period of 2^{32}, we would only allow for about 400 replications in a panel data simulation study with $NT = 1000$ and 10 variables. So, some care is needed when using SAS and RATS RNG. Second, even if the fact to pass the DIEHARD test (or other ones) does not prove that a RNG is a performing one, but failure to pass several tests proves that it should not be used. These two issues are important since the quality of a uniform RNG is crucial for the quality of a non uniform one and, so determines in part the validity of the simulation results. On this last ground, users may be confident with the 8 software we reviewed.

Acknowledgments I am grateful to Aptech System Inc. (GAUSS), Econometric Software Inc. (LIMDEP), Estima (RATS), Quantitative Micro Software (EViews), Stata Corp., Timberlake Consultants (Ox), TSP International, SAS Institute, for having provided me freely a copy for review purposes.

I sent a prior draft of the review to each of the software editors for checking. Any remaining errors are my own. I am grateful to C. Cummins (TSP), W. Greene (LIMDEP, NLOGIT), G. Liang (EViews), T. Doan (RATS), R. Schoenberg (GAUSS), D. M. Drukker (Stata) for helpful comments and advice. I would like to thank M. Harris for helpful comments on an earlier draft.

Appendix: General Product Information

Table 26.A.1 Software general informations

Name	Platform/OS[a]	Student Version	Prices[b]	Web Site
EViews 6 (2007)	Win98+	Yes[l] (v.3, $40)	$495[k] (V. 6)	http://www.eviews.com
GAUSS 8 (2007)	Windows (32-bit) Unix, Linux	No	$1000	http://www.aptech.com
LIMDEP[c] 9 (2007)	Windows (32-bit)	Yes[d] (free)	LIMDEP $595 Nlogit $795	http://www.limdep.com
Ox[e] 4.1 (2007)	Windows Mac Unix, Linux	Yes (85£)[f]	£250	http://www.doornik.com
RATS 6.35 (2007)	Windows, Mac Unix, Linux	Yes ($300)[g]	$500	http://www.tspintl.com
TSP 5 (2005)	Windows, Mac Unix, Vax-Vms	Yes ($100)[h]	$500[e]	http://www.estima.com
SAS 9.1 (2003)	All except Macintosh	Yes[i]	[i]	http://www.sas.com
Stata 10 (2007)	Windows, Mac Unix, Linux	No[j]	$985 (v. SE 10)[k]	http://www.stata.com

[a] Win98+ = Windows 98 and later, WinNT+ = Windows NT, 2000, XP.
[b] Academic, Windows version, one copy.
[c] With NLOGIT.
[d] EA-Limdep: http://www.prenhall.com/greene/ealimdep.htm
[e] With GiveWin.
[f] OxConsole (DOS/Unix) free.
[g] $60 with a limit of 6000 obs.
[h] With a limit of 16000 numbers. $75 without GiveWin.
[i] according to country.
[j] Small Stata 8, $400.
[k] With complete documentation set.

References

Allison, P. D. (1999). *Logistic Regression Using the SAS System: Theory and Application*. John Wiley & Sons, New York.

Anderson, T. W. and Hsiao, C. (1981). Estimation of Dynamic Models with Errors components, *Journal of the American Statistical Association*, 76:598–606.

Arellano, M. and Bond, S. (1991). Some Tests of specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, 58:277–297.

Arellano, M. and Bond, S. (1998). Dynamic panel Data Estimation Using DPD98 for GAUSS: A Guide for Users. *Mimeo*.

Barrachina, R. and Sanchis Llopis, J. A. (2002). Stata 7.0 para windows. *Revista de Economia Aplicada*, 10:163–174.

Cribari-Neto, F. (1997). Econometric Programming Environments: GAUSS, Ox and S-Plus. *Journal of Applied Econometrics,* 12:77–89.

Cribari-Neto, F. (1999). C for Econometricians, *Computational Economics.* 14:135–149**.**

Cribari-Neto, F. and Zarkos, S. G. (2003). Econometric and Statistical Computing Using Ox. *Computational Economics,* 21:277–295**.**

Doornik, J. A. (2001). *Object-Oriented Matrix Programming using Ox.* 4^{th} edition. Timberlake Consultants Press, London.

Doornik, J. A. (2004). *Ox Appendices. Mimeo.*

Doornik, J. A., Arellano, M. and Bond, S. (2002). Panel data estimation using DPD for Ox. *Mimeo*.

Doornik, J. A., Draisma G. and Ooms M. (2001). *Introduction to Ox*. Timberlake Consultants Press, London.

Eddelbüttel, D. (1996). Object-Oriented Econometrics: Matrix Programming in C++ Using GCC and NEWMAT. *Journal of Applied Econometrics*, 11:199–209.

Eklöf, M. and Weeks, M. (2004*a*). Estimation of Discrete Choice Models Using DCM for Ox. *Mimeo*.

Eklöf, M. and Weeks, M. (2004*b*). Discrete Choice Models (DCM): An Object-Oriented Package for Ox. *Mimeo*.

Fry, T. R. (1996). LIMDEP 7.0. *Journal of Economic Surveys,* 10:461–465.

Gill, L., Guthrie, W., Lagergren, E., and Rogers, J. (2001). Statistical Reference Datasets (StRD) for Assessing the Numerical Accuracy of Statistical Software, *Proceedings of the Joint Statistical Meetings*, Anaheim.

Gould, W., Pitblado, J. and Sribney, S. (2003). *Maximum Likelihood Estimation with Stata.* 2^{nd} edition. Stata Press College Station.

Heywood, G. and Rebelo, I. (1996). Time series analysis of cross-sections: A software guide. *The Economic Journal,* 106:271–275.

Holly, S. and Turner, P. (1998). Programming Facilities in EViews. *The Economic Journal,* 108:1632–1639.

Hsiao, C. (1986). *Analysis of Panel Data.* Cambridge University Press Cambridge.

Keane, M. P. and Wolpin, K. I. (1997). The Career Decisions of Young Men. *Journal of Political Economy*, 105:473–522.

Kenc, T. and Orszag, J. M. (1997). Ox: an object-oriented matrix language. *The Economic Journal*, 107:256–259.

Kerninghan, B. W. and Ritchie, D. M. (1988). *The C Programming Language*. 2nd edition. Prentice Hall New Jersey.

Kolenikov, S. (2001). Review of Stata 7. *Journal of Applied Econometrics*, 16:637–646.

Küsters, U. and Steffen, J. P. (1996). Matrix Programming Languages for Statistical Computing: A Detailed Comparison of GAUSS, Matlab, and Ox. *mimeo,* Katholischen Universität Eichstätt.

Laurent, S. and Urbain, J.-P. (2004). Bridging the Gap Between Ox and GAUSS using OxGAUSS. *mimeo*.

Lee, M.-T. (2002). *Panel Data Econometrics*, Academic Press, San Diego.

Lilien, D. M. (2000). Econometric Software Reliability and Nonlinear Estimation in EViews: Comment. *Journal of Applied Econometrics,* 15:107–111.

Lubrano, M. (1998). TSP 4.4: Old but Powerfull Econometrics Software. *The Economic Journal,* 108:1621–1627.
McCullough, B. D. (1997). A Review of RATS v4.2: Benchmarking Numerical Accuracy. *Journal of Applied Econometrics,* 12:181–190.
McCullough, B. D. (1998). Assessing the Reliability of Statistical Software: Part I. *The American Statistician*, 52:358–366.
McCullough, B. D. (1999*a*). Assessing the Reliability of Statistical Software: Part II. *The American Statistician*, 53:149–159.
McCullough, B. D. (1999*b*). Econometric Software Reliability: EViews, LIMDEP, SHAZAM and TSP. *Journal of Applied Econometrics,* 14:191–202.
McCullough, B. D. and Vinod, H. D. (1999). The Numerical Reliability of Econometrics Software. *Journal of Economic Literature*, 37:633–665.
McKenzie, C. R. (2000). LIMDEP, Version 7, *The Economic Journal,* 110:F455–F461.
McKenzie, C. R. and Takaoka, S. (2003). 2002: A LIMDEP Odyssey. *Journal of Applied Econometrics*, 18:241–247.
Merino, F. (1998). TSP 4.4 for non time series. *The Economic Journal,* 108:1627–1632.
Nerlove, M. (1971). Further Evidence on the Estimation of Dynamic Economic Relations From a Time Series of Cross Sections. *Econometrica*, 39:359–381.
Nerlove, M. (1998). GAUSS and Matlab: Comparison and Evaluation with a Supplementary Discussion of MAPLE and MATHEMATICA. *mimeo*.
Nerlove, M. (2001). On the Numerical Accuracy of Mathematica 4.1 for Doing Ordinary Least-Squares Regressions. *mimeo*.
Nerlove, M. (2002). *Essays in Panel Data Econometrics.* Cambridge University Press Cambridge.
Noble, N. R. (1999). Software Reviews: EViews 3.0. *International Journal of Forecasting*, 15:222–224.
Podovinsky, J. M. (1999). Ox 2.10: Beast of Burden or Object of Desire? *Journal of Economic Surveys*, 13:491–502.
Roberts, J. (1995). Econometrics Views Version 1.0. *The Economic Journal*, 105:1048–1055.
Romeu, A. (2004). ExpEnd: GAUSS Code for Panel Count-data Models. *Journal of Applied Econometrics*, 19(3):429–434.
Silk, J. (1997). TSP 4.4: A Review, *Journal of Applied Econometrics*, 12:445–453.
Sparks, G. R. (1997). Econometric Views 2.0, *Journal of Economic Surveys*, 11:107–113.
Stata Corp., (2003). *Cross-Sectional Time-Series.* Stata Press, Collage Station.
Steinhaus, S. (2002). Comparison of Mathematical Programs for Data Analysis. *mimeo*.
Stroustrup, B. (1992). *The C++ Programming Language.* Addison-Wesley, Murray Hill NJ.
Train, K. (2003). *Discrete Choice Methods with Simulation.* Cambridge University Press Cambridge.
Vinod, H. D. (2000). Review of GAUSS for Windows, Including its Numerical Accuracy. *Journal of Applied Econometrics*, 15:211–220.
Viton, P. A. (2003). Running GAUSS Programs Under Ox3. *Mimeo.*
Windmeijer, F.(2000). A finite sample correction for the variance of linear two-step GMM estimators. IFS working paper W00/19.
Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data.* MIT Press.
Wright, R. E. (1996). LIMDEP: Limited Dependant Variables Models, version 7.0. *The Economic Journal*, 106:1458–1460.

Advanced Studies in Theoretical and Applied Econometrics

1. J.H.P. Paelinck (ed.): *Qualitative and Quantitative Mathematical Economics*. 1982 ISBN 90-247-2623-9
2. J.P. Ancot (ed.): *Analysing the Structure of Econometric Models*. 1984 ISBN 90-247-2894-0
3. A.J. Hughes Hallet (ed.): *Applied Decision Analysis and Economic Behaviour*. 1984 ISBN 90-247-2968-8
4. J.K. Sengupta: *Information and Efficiency in Economic Decision*. 1985 ISBN 90-247-3072-4
5. P. Artus and O. Guvenen (eds.), in collaboration with F. Gagey: *International Macroeconomic Modelling for Policy Decisions*. 1986 ISBN 90-247-3201-8
6. M.J. Vilares: *Structural Change in Macroeconomic Models*. Theory and Estimation. 1986 ISBN 90-247-3277-8
7. C. Carraro and D. Sartore (eds.): *Development of Control Theory for Economic Analysis*. 1987 ISBN 90-247-3345-6
8. D.P. Broer: *Neoclassical Theory and Empirical Models of Aggregate Firm Behaviour*. 1987 ISBN 90-247-3412-6
9. A. Italianer: *Theory and Practice of International Trade Linkage Models*. 1986 ISBN 90-247-3407-X
10. D.A. Kendrick: *Feedback*. A New Framework for Macroeconomic Policy. 1988 ISBN 90-247-3593-9; Pb: 90-247-3650-1
11. J.K. Sengupta and G.K. Kadekodi (eds.): *Econometrics of Planning and Efficiency*. 1988 ISBN 90-247-3602-1
12. D.A. Griffith: *Advanced Spatial Statistics*. Special Topics in the Exploration of Quantitative Spatial Data Series. 1988 ISBN 90-247-3627-7
13. O. Guvenen (ed.): *International Commodity Market Models and Policy Analysis*. 1988 ISBN 90-247-3768-0
14. G. Arbia: *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. 1989 ISBN 0-7923-0284-2
15. B. Raj (ed.): *Advances in Econometrics and Modelling*. 1989 ISBN 0-7923-0299-0
16. A. Aznar Grasa: *Econometric Model Selection*. A New Approach. 1989 ISBN 0-7923-0321-0
17. L.R. Klein and J. Marquez (eds.): *Economics in Theory and Practice*. An Eclectic Approach. Essays in Honor of F. G. Adams. 1989 ISBN 0-7923-0410-1
18. D.A. Kendrick: *Models for Analyzing Comparative Advantage*. 1990 ISBN 0-7923-0528-0
19. P. Artus and Y. Barroux (eds.): *Monetary Policy*. A Theoretical and Econometric Approach. 1990 ISBN 0-7923-0626-0
20. G. Duru and J.H.P. Paelinck (eds.): *Econometrics of Health Care*. 1990 ISBN 0-7923-0766-6
21. L. Phlips (ed.): *Commodity, Futures and Financial Markets*. 1991 ISBN 0-7923-1043-8

Advanced Studies in Theoretical and Applied Econometrics

22. H.M. Amman, D.A. Belsley and L.F. Pau (eds.): *Computational Economics and Econometrics*. 1992 ISBN 0-7923-1287-2
23. B. Raj and J. Koerts (eds.): *Henri Theil's Contributions to Economics and Econometrics*. Vol. I: Econometric Theory and Methodology. 1992 ISBN 978-0-7923-1548-3
24. B. Raj and J. Koerts (eds.): *Henri Theil's Contributions to Economics and Econometrics*. Vol. II: Consumer Demand Analysis and Information Theory. 1992 ISBN 978-0-7923-1665-7
25. B. Raj and J. Koerts (eds.): *Henri Theil's Contributions to Economics and Econometrics*. Vol. III: Economic Policy and Forecasts, and Management Science. 1992 ISBN 978-0-7923-1664-0
    Set (23–25) ISBN 978-0-7923-1666-4
26. P. Fisher: *Rational Expectations in Macroeconomic Models*. 1992 ISBN 978-0-7923-1903-0
27. L. Phlips and L.D. Taylor (eds.): *Aggregation, Consumption and Trade*. Essays in Honor of H.S. Houthakker. 1992 ISBN 978-0-7923-2001-2
28. L. Matyas and P. Sevestre (eds.): *The Econometrics of Panel Data*. Handbook of Theory and Applications. 1992 ISBN 978-0-7923-2043-2
29. S. Selvanathan: *A System-Wide Analysis of International Consumption Patterns*. 1993 ISBN 978-0-7923-2344-0
30. H. Theil in association with D. Chen, K. Clements and C. Moss: *Studies in Global Econometrics*. 1996 ISBN 978-0-7923-3660-0
31. P.J. Kehoe and T.J. Kehoe (eds.): *Modeling North American Economic Integration*. 1995 ISBN 978-0-7923-3751-5
32. C. Wells: *The Kalman Filter in Finance*. 1996 ISBN 978-0-7923-3771-3
33. L. Matyas and P. Sevestre (eds.): *The Econometrics of Panel Data*. A Handbook of the Theory with Applications. 2nd Revised ed. 1996 ISBN 978-0-7923-3787-4
34. B. Assarsson, D. Edgerton, A Hummelmose, I Laurila, K. Rickertson and P.H. Vale: *The Econometrics of Demand Systems*. With Applications to Food Demand in the North Countries. 1996 ISBN 978-0-7923-4106-2
35. D.A. Griffith, C.G. Amrhein and J-M. Huriot (eds.): *Econometric Advances in Spatial Modelling and Methodology*. Essays in Honour of Jean Paelinck. 1998 ISBN 978-0-7923-4915-0
36. R.D.H. Heijmans, D.S. G. Pollock and A. Satorra (eds.): *Innovations in Multivariate Statistical Analysis*. 2000 ISBN 978-0-7923-8636-0
37. R. MacDonald and I. Marsh: *Exchange Rate Modelling*. 2000 ISBN 978-0-7923-8668-1
38. L. Bauwens and P. Giot: *Econometric Modelling of Stock Market Intraday Activity*. 2001 ISBN 978-0-7923-7424-4
39. J. Marquez: *Estimating Trade Elasticities*. 2002 ISBN 978-1-4020-7159-1
40. R. Bhar and S. Hamori: *Hidden Markov Models*. Applications to Financial Economics. 2004 ISBN 978-1-4020-7899-6
41. P. Tryfos: *The Measurement of Economic Relationships*. 2004 ISBN 978-1-4020-2838-0

42. M.O. Haque: *Income Elasticity and Economic Development*. Methods and Applications. 2005 ISBN 978-0-387-24292-7
43. I.F. Razafimahefa and S. Hamori: *International Competitiveness in Africa*. Policy Implications in the Sub-Saharan Region. 2007 ISBN 978-3-540-68920-1
44. C. G. Renfro: *The Practice of Econometrics Theory*. 2008 *Forthcoming* ISBN 978-3-540-75570-8
46. L. Mátyás and P. Sevestre: *The Econometrics of Panel Data.* Fundamentals and Recent Developments in Theory and Practice. 3rd ed. 2008 ISBN 978-3-540-75889-1