Chapter 3
Multivariate and Multilevel Longitudinal Analysis

Nicholas T. Longford

Abstract This chapter presents a review of perspectives and methods for analysis of longitudinal data on several related variables. A connection is made with multilevel analysis in which the longitudinal and multivariate dimensions of the data can naturally be subsumed. With the focus on large-scale longitudinal studies of human subjects who are in general disinterested in and not highly motivated by the agenda of the study, methods for dealing with nonresponse are an essential addendum to the analytical equipment.

3.1 Introduction

Modern practice of data collection from human subjects is highly aware of the costs and difficulties in retaining survey respondents, especially in longitudinal studies in which survey subjects are to be contacted on several occasions, sometimes over a long period of time. One reaction to these pressures is to collect more complete information from complying subjects, so that the resulting data would be well suited for a wider analytical agenda within the remit of the survey. In particular, it would enable us to study the associations of several variables, and how these associations are altered over time.

In this perspective, it is more appropriate to consider as an elementary data item the value of a vector $\mathbf{X}^{(t)}$ observed on a (single) occasion t. Any one component of $\mathbf{X}^{(t)}$ offers little information without the other components of $\mathbf{X}^{(t)}$. However, the vector $\mathbf{X}^{(t)}$ offers only a snapshot of a social, economic or epidemiological development in the studied population so, for any single t, $\mathbf{X}^{(t)}$ is also a much poorer source of information than the sequence $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(T)}$. Such a sequence can be presented as a random matrix, and data for a random sample from the population as a

Nicholas T. Longford
SNTL, Barcelona, Spain
e-mail: NTL@sntl.co.uk

K. van Montfort et al., *Longitudinal Research with Latent Variables*,
DOI 10.1007/978-3-642-11760-2_3,

three-dimensional array $\mathbf{X}$ composed of vectors $\mathbf{x}_i^{(t)}$ for subjects $i = 1,\ldots,n$ at time points $t = 1,\ldots,T$.

We assume that the goal of an analysis is inference about a particular finite but large population $\mathscr{P}$, and that this population is represented by a sample $\mathscr{S}$ drawn from $\mathscr{P}$ by a simple random sampling design. We assume that the time points $1,\ldots,T$, at which the values of a vector of variables $\mathbf{X}$ are observed, are selected noninformatively, without regard for any of the values $\mathbf{X}$ of the units in the sample.

We have two perspectives which lead to diverging approaches to inference. In the *sampling-design based perspective*, there is a finite set of units $1,\ldots,N$ with fixed (unchanging) values of $\mathbf{X}^{(t)}$ for every time point t. In a replication of the study, a different set of units would be selected into the sample, but if a unit i happened to be selected in both samples, its values of $\mathbf{X}^{(t)}$ would be the same in the two replications. In this view, sampling is the only source of variation, and the sampling design provides its complete description.

In the *model-based perspective*, the values of $\mathbf{X}^{(t)}$, $t = 1,\ldots,T$, are generated by a particular stochastic process, the definition of which (or, in most practical settings, an approximation to it) is the analyst's responsibility. Inferences are made assuming this model, but the analysis is accompanied by a careful diagnosis that searches for contradictions of the data with the assumptions made. This approach is much more common nowadays because it is more flexible, with a greater variety of software tools that have the necessary elements for its implementation.

The two perspectives are not completely separated. Dealing with nonresponse is a notable concern that they have in common. Even in the sampling-design based perspective, a model has to be posited for how the missing data are related to the recorded data (Little and Rubin, 2002); without a model the analysis would be at a dead end. In contrast, the model-based perspective ignores all the units with empty records (no data available); in many analyses no information is available about the units that were selected into the sample but nothing was recorded about them. The perspective is, however, concerned about making use of the information in incomplete records for which some, but not all, values are recorded. The concern about good representation of a population often appears out of place because no reference population is defined, or the model is specified in such a way that it implies or generates an impression of universality; that, within some reasonable bounds, it applies to *any* population.

Our view is that this perspective is constructive but not valid. We qualify this view by adding that we do not regard model validity as an imperative for a respectable analysis. We illustrate this on a simple example of a growth model

$$\mathbf{y}_i = \mathbf{Z}_i\beta + \varepsilon_i\,, \tag{3.1}$$

where $\mathbf{y}_i$ are $T \times 1$ (column) vectors of outcomes for units $i = 1,\ldots,n$, $\mathbf{Z}_i$ is the regression matrix for unit i, β is the vector of regression parameters, and ε_i are a random sample from a centred multivariate normal distribution, $\mathscr{N}(\mathbf{0},\Sigma)$. We may regard $\mu_i = \mathbf{Z}_i\beta$ as the growth for a typical unit, but deviations from μ_i, unless they are extreme, cannot be regarded as anything untypical. The vector of deviations ε_i

does not represent any errors, because deviation from μ_i is not necessarily a sign of anything that has gone wrong. Usually incorrect is the central assumption of the functional form of μ_i — an unavoidable 'error' committed by the analyst against the environment (nature) in which units are exposed to a multitude of influences, many of them continual, the impact of which defies both our understanding and any neat algebraic summarisation. The choice of the variables in $\mathbf{Z}$ is governed by analytical pragmatism, attempting to capture the most important features of the studied phenomenon. With more extensive data (more observed units), we can capture finer detail and include more variables in $\mathbf{Z}$. When more variables are recorded, we have a wider choice of variables in $\mathbf{Z}$. Validity of a model, defined as a collection of distributions according to one of which the data is generated, is an unattainable goal. Its pragmatic reduction is a model that the data appear not to contradict, as assessed by various model-diagnostic procedures.

In theory and reality, there is a single valid model (the process); in practice, we improvise with the information we possess, and the intermediate goal of variable selection has different targets depending on the extent of the information, ignoring the fact that there is only *one* valid model. The pretense that the model we have selected is the valid model is a common logical inconsistency that does considerable harm to the integrity of the statistical practice. Attempts at addressing this problem (Draper, 1995; Chatfield, 1995; and Longford, 2007) have been largely ignored because of the complexity involved. They entail taking into account the model uncertainty, acknowledging that the model-selection process is also subject to sampling variation.

The model in (3.1) ascribes a different status to the covariates in $\mathbf{Z}$ than to the outcomes in $\mathbf{y}$, even when there is no distinction in the way their values are collected in a survey. Both $\mathbf{y}$ and $\mathbf{Z}$ are attributes of the members of the population that are not amenable to any control, unlike a treatment assigned by randomisation in an experimental study. In particular, any causal inference is highly problematic when $\mathbf{Z}$ is observed just as passively as $\mathbf{y}$, without exercising any influence (control) over its values. The regression in (3.1), summarised by the vector of parameters β, is a comparison of subpopulations (strata) defined by the values of $\mathbf{Z}$, and it offers no basis for statements about manipulation — what would happen if a particular unit had a different value of $\mathbf{Z}$. There would be an answer, in principle, if the valid data-generating model were known. In practice, such a model is not known and the recorded variables are usually a small subset of the variables that would have a role in such an ideal model.

In the modelling perspective, longitudinal analysis combines aspects of multivariate and multilevel analyses. It is multivariate, because one or several variables are observed on several occasions, and the study of the associations of these (time-specific) versions of the variable(s) is of obvious interest. It is multilevel, because the observations on a subject at the time points are naturally clustered, and the subjects may be further clustered within families, areas (locations), schools, businesses and similar organisations. The purpose of this chapter is to elaborate these links and perspectives, with an emphasis on taking advantage of their strengths in responding to the various complexities encountered in the analysis of longitudinal data.

The next section introduces the univariate longitudinal setting and the following section discusses nonresponse. Section 3.4 extends the models for multivariate outcomes. Section 3.5 discusses modelling of univariate outcomes in greater detail, studying dependence across time and variance heterogeneity. Section 3.6 deals with multivariate versions of these models. Computational issues, model fitting and graphical presentation, are addressed in Section 3.7. The chapter is concluded with a discussion.

3.2 Inferential Targets

Assuming that the values of the vector of outcomes $\mathbf{x}^{(t)}$ are well defined for any time-point $t \in (0,T)$, or beyond, we may associate each member j of $\mathscr{P}$ with a multivariate function $F_j(t)$ of time. This function, describing the growth, evolution or development, is a relevant target of inference. Inference about its behaviour in the near future amounts to extrapolation, but we can learn from its behaviour in the past, assuming some form of stationarity. The observations $\mathbf{x}_j^{(t)}$ at time points $t = 1,\ldots,T$ inform about F_j only partially. If all the subjects in the sample are observed in a regimented fashion, at time points $t = 1,\ldots,T$, then we have no information about the behaviour of F_j between any two (integer) time points. This suggests that we may learn more by implementing designs with unevenly set time points t. The vectors of outcomes may have uneven lengths, and the time points for a unit need not be distributed evenly. However, the choice of the time points t has to be noninformative for every unit, independently of the functions F_j. This is ensured when the time points are set by design. When the observational units (subjects) volunteer to provide the information, (e.g., as patients or customers), or become data donors opportunistically e. g., by being met at a railway station or a shopping centre, we have to be concerned about the good representation of the sample, as well as by the non-ignorable nature of the time-selection process.

The model in (3.1) has no straightforward adaptation for unevenly distributed time points. For each unit i we posit a model

$$y_{ih} = f_i(t_{ih}) + \varepsilon_{ih}\,,$$

where t_{ih} is the time at the observation h of unit i and ε_{ih} are a random sample from a (univariate) centred normal distribution, $\mathscr{N}(0,\sigma^2)$. We may specify a separate model for the variance σ^2, relating it to time t. The unit-specific functions f_i may involve some coefficients ξ_i, for which another model would be defined, linking the units to vectors ξ_i:

$$\xi_i = \nu + \delta_i\,, \tag{3.2}$$

where δ_i is a random sample from a multivariate distribution. Instead of ν we may have a model that relates the expectations $\mathrm{E}(\xi)$ to a (linear) function of some covariates defined for the units. The decomposition in (3.2) connects the unit-specific

functions f_i and enables us to describe the population of units in terms of a typical unit given by the parameter vector ν and unit-level variation described by the distribution of δ_i. The link between f_i and ξ_i need not be linear, and so the function f that corresponds to ν is, in general, not a population average of the functions f_i.

3.3 Incompleteness

By complete data we understand a valid entry for every data item that was intended to be collected by the design (protocol). A typical protocol calls for collecting a rectangular dataset, a list of variables recorded at each of a set of time points for every unit in the sample. Incompleteness, broadly interpreted as failure to adhere to the design, is common especially when the units are human subjects for whom the interview and measurement (elicitation) process are an unwelcome distraction. A record comprising entirely of missing values (unit nonresponse) or lost in the process of transfer from the interviewer (data collector) to the (secondary) analyst through the database constructor, may be dropped from the analysis. If no trace is left after such records in the database the analyst knows nothing about their existence.

A record comprises subrecords for the time points, and any of these subrecords may be missing (time-point or *wave* nonresponse). Unless the analyst is aware, or infers from the patterns in the data, that the design called for the collection of a rectangular dataset, the dataset can be subjected to an analysis as if it were complete. Similarly, a subrecord may be empty or incomplete, involving item nonresponse. The design, however, is important. Pretending that the incomplete dataset is complete results in invalid inferences — inappropriate claims of unbiasedness and efficiency.

Even if the design did not call for a rectangular dataset, we may pose the problem of the analysis as involving missing values, values the addition of which would make the dataset rectangular and amenable to a relatively simple analysis. Of course, this approach is not practical when a large fraction of the values in the hypothetical rectangular dataset are missing (and have to be imputed) and the pattern of nonresponse is varied. When practical, this approach is relatively simple to implement because we are privy to the details of the nonresponse process.

3.4 From Univariate to Structured Multivariate Data

We develop models for multivariate longitudinal data within a more general framework of multivariate structured outcomes from univariate models and data by adding dimensions. We use the term *dimension* similarly to the term *factor* in ordinary regression (and the software `GLIM`; Francis, Green and Payne, 1993). Thus the various outcomes recorded on an occasion are a dimension, and the times of observation are another dimension. We refer to the outcomes as *components* of the vector of

outcomes, although the components themselves can be multivariate; for example, a row of the timepoint-by-variable matrix of outcomes of a subject is a component.

For a univariate outcome y we consider linear regression on some covariates $\mathbf{x}$:

$$y = \mathbf{x}\beta + \varepsilon\,,$$

with the usual assumptions of independence, normality and homoscedasticity. We address departures from normality by generalized linear models, for which an alternative distributional assumption is required, together with a link function that relates the underlying linear predictors to the conditional expectations of the outcomes, $\mathrm{E}(y\,|\,\mathbf{x})$.

The structure of clustering, of sets of units having more similar values of the outcome y than the units in general, is introduced by assuming that the units within clusters are correlated. The simplest way of doing this is by the *compound symmetry* model, in which

$$\mathbf{y}_j = \mathbf{X}_j\beta + \delta_j + \varepsilon_j\,, \tag{3.3}$$

where $\mathbf{y}_j = (y_{1j},\ldots,y_{n_j j})^\top$ is the vector of outcomes in cluster j, $\mathbf{X}_j$ is the regression matrix for these units, composed of the rows $\mathbf{x}_{ij}$; δ_j, $j = 1,\ldots,n$, are a random sample from $\mathscr{N}(0,\sigma_{\mathrm{B}}^2)$; the $n = n_1 + \cdots + n_m$ elements of ε_j are a random sample from $\mathscr{N}(0,\sigma^2)$; and the two random samples are independent. The within-cluster correlation $\rho = \sigma^2/(\sigma^2+\sigma_{\mathrm{B}}^2)$ summarises the relative similarity of the units within clusters.

For observational (elementary) units within clusters we can distinguish between variables that are defined for the units (elements) and for the clusters. The latter variables are expanded for the elements so that all units within a cluster have the same value as their cluster. Variables defined for units could, in principle, have values that are constant within clusters. At the other extreme, the values could have identical means, or even identical distributions within the clusters. Such variables are called *balanced* with respect to clustering. As a convention, we include the intercept, represented by the vector of unities $\mathbf{1}$, among the balanced variables. For the vectors of covariates $\mathbf{x}_{ij}$ we have the following decomposition of the matrix of crossproducts:

$$\mathbf{X}^\top\mathbf{X} = \mathbf{B} + \mathbf{W}\,, \tag{3.4}$$

where

$$\mathbf{B} = \sum_{j=1}^{m} n_j\,(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^\top\,(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})$$

$$\mathbf{W} = \sum_{j=1}^{m}\sum_{i=1}^{n_j} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)^\top\,(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)$$

and $\bar{\mathbf{x}}_j$ is the sample mean within cluster j and $\bar{\mathbf{x}}$ the overall sample mean. Balanced variables contribute only to $\mathbf{W}$ and cluster-level variables only to $\mathbf{B}$.

We have to draw a distinction between the sample and population versions of the summaries **B** and **W**, as well as any other quantities we define later. Although the models we consider have fixed values of the covariates, **X**, in a typical sampling process applied in a (human) population the values of **X** are random. That is, in a hypothetical replication of the survey, a different matrix **X** would be realised. In the (sampling) design-based perspective, the values of **X** and **y** are fixed in the population, and the sampling process is the sole source of variation. That is, if a subject happened to be included in the sample in two replications, his or her values of **x** would be the same, and he or she would be in the same cluster.

The design-based perspective has in the past been regarded as not constructive, and the inferential effort in many areas has drifted toward model-based approaches. However, there are areas where the balance is being restored. For example, the potential outcomes framework for observational studies (Holland, 1986; Rubin, 2005) shifts the focus from the association of **X** and **y** to the analysis of the (treatment) assignment process. This analysis is model-based, but it is only an intermediary to the substantive analysis which follows, and which is simple, related to the analysis in an experimental setting, and has more in common with the design-based paradigm.

In the model-based paradigm, the similarity of the units within clusters can be interpreted in terms of differing within-cluster associations of **X** and **y**. The model in (3.3) corresponds to parallel within-cluster regressions, which have identical regression slopes, but different intercepts $\beta_0 + \delta_j$. This characterisation uncovers its relative rigidity. Much greater flexibility is attained by allowing some (or all) regression slopes to vary from cluster to cluster. The within-cluster slope for a variable that is constant within clusters is not identified. Therefore it is meaningful to consider varying slopes only with respect to variables defined for the elements. We split the covariates into the two groups, $\mathbf{X} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$, where $\mathbf{X}^{(1)}$ are defined for elements and $\mathbf{X}^{(2)}$ for clusters; we assume that none of the variables in $\mathbf{X}^{(1)}$ is constant within clusters. Then the compound symmetry model is

$$\mathbf{y}_j = \mathbf{X}_j^{(1)}\beta^{(1)} + \mathbf{X}_j^{(2)}\beta^{(2)} + \delta_j + \varepsilon_j .$$

Its obvious generalisation is

$$\mathbf{y}_j = \mathbf{X}_j^{(1)}\beta^{(1)} + \mathbf{X}_j^{(2)}\beta^{(2)} + \mathbf{X}_j^{(1)}\delta_j + \varepsilon , \tag{3.5}$$

where δ_j is a random sample from a centred multivariate normal distribution, $\mathcal{N}(\mathbf{0}, \Sigma_{\mathrm{B}})$. We have to extend the definition of the multivariate normal distribution to singular (degenerate) distributions for which Σ_{B} is singular. Let p_h be the number of covariates (columns) in $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. Then Σ_{B} is a $p_2 \times p_2$ variance matrix. In Σ_{B}, it is meaningful to constrain some variances to zero. This corresponds to constant within-cluster slopes with respect to the corresponding covariates. When a variance is constrained to zero, then so are all the covariances in the same row and column. We have to obey the rules of invariance with respect to linear transformations (Longford, 2007, Chapter 9), which dictate that the intercept should be associated with a variance to be estimated so long as any other covariate is. A categorical

variable with K distinct labels is represented among covariates by $K-1$ indicator variables. When such a variable is defined for elements, the invariance rules imply that either all the $K-1$ variables are associated with variances to be estimated, or none are. After all, the values of the indicator variables are contingent on the choice of the reference, which in most cases is arbitrary or opportunistic.

3.5 Univariate Observations at Time Points

In longitudinal analysis, each unit j is observed at a (finite) sequence of time points

$$\mathbf{t}_j = \left(t_j^{(1)}, t_j^{(2)}, \ldots, t_j^{(n_j)}\right)^\top .$$

When all units are observed at the same set of time points, $\mathbf{t}_j \equiv \mathbf{t}$, the outcomes form a sample from a multivariate (normal) distribution, so that

$$\mathbf{y}_j \sim \mathcal{N}(\mu, \Sigma) .$$

Structures can be imposed on the vector μ and variance matrix Σ, such as linear growth and compound symmetry, but these are useful only when the number of time points, p, is large, so that a linear function, represented by two parameters, or a quadratic function, by three, provides a much more compact description for the growth (development, expansion, decay, or the like) than the components of μ. The unstructured vector μ is 'always correct', but may be ineffective, in that it restricts the inferences that can be made to the specific time points.

In contrast, a functional expression for μ is in general incorrect, but the bias it entails may be offset by the reduced sampling variance associated with its estimation. The multivariate perspective is inflexible — it cannot be adapted for inferences about other time points, by inter- or extrapolation. The functional perspective caters for such inferences by prediction, although the issues of correctness, and its scope being limited to the particular context, delimit its application, especially for extrapolation.

Without a structure imposed on μ, the design has to ensure that the time points in $\mathbf{t}$ are the ones for which inference is desired. With a structure on μ, the design has to ensure that the function underlying the expectations μ, $\mu(t)$, can be estimated with desired precision and predictions based on it have sufficient quality.

Similar comments can be made about specifying Σ. Without a structure imposed on Σ, each covariance in Σ is a unique quantity, although in most contexts we can reasonably assume that greater distance of the time points is associated with lower correlation. For the variances in Σ, a reasonable assumption may be that they are constant or increasing with the distance in time, but a function underlying them amounts to an assumption highly contingent on (the choice and coding of) the time points $\mathbf{t}$.

A parametric structure can be imposed on Σ or on its inverse, Σ^{-1}, called the *concentration matrix* or, in principle, on any transformation of Σ. Working with Σ^{-1} is particularly attractive when considering the Markov property of conditional independence of any two observations given an observation that separates them:

$$\left(y_{t_1} \,|\, y_{t_2}, y_{t_3}\right) \sim \left(y_{t_1} \,|\, y_{t_2}\right),$$

for the outcomes at any time points $t_1 < t_2 < t_3$. The corresponding matrix Σ^{-1} is tridiagonal:

$$\left\{\Sigma^{-1}\right\}_{kh} = 0$$

whenever $|k-h| \geq 2$. When the number of time points is greater, this constraint may be considered also for $|k-h| \geq 3$ or 4; that is, entries outside the diagonal strip of Σ^{-1} of the given width vanish.

The parameters in μ and Σ are indivisible in the following sense. When no structure is imposed on μ a structure should not be imposed on Σ either. Imposing a structure on μ is simpler than on Σ, because it is a unidimensional object. Therefore a structure may be imposed on μ, but not on Σ, but this is mainly a pragmatic matter reflecting our inability or lack of confidence about specifying suitable submodels.

Observations of the outcomes Y in a longitudinal analysis may be accompanied by the values of covariates. These may be defined for the subjects and for the (elementary) observations. Adjustment for the variables defined for subjects may be made by multivariate regression:

$$\mathbf{y}_j = \mathbf{x}_j \mathbf{B} + \gamma_j, \tag{3.6}$$

where $\mathbf{y}_j$ is the vector of outcomes for subject j, $\mathbf{x}_j$ the vector of values of the covariates, $\mathbf{B}$ the matrix of regression parameters, and the deviations γ_j are a random sample from a multivariate normal distribution, $\mathcal{N}(\mathbf{0}, \Gamma)$.

Covariates that are specific to time points cannot be accommodated in the model in (3.6) because the vector $\mathbf{y}_j$ is treated like a single unit. The problem does not arise with hierarchical models in which observations and subjects form separate levels of nesting:

$$y_{ij} = \mathbf{x}_j^{(1)} \beta^{(1)} + \mathbf{x}_j^{(2)} \beta^{(1)} + \mathbf{x}_j^{(1)} \delta_j + \varepsilon_{ij}, \tag{3.7}$$

with assumptions similar to those in (3.5). The variables in $\mathbf{x}_j^{(1)}$ are defined for the occasions and those in $\mathbf{x}_j^{(2)}$ for subjects. In the variance matrix $\Sigma_{\mathrm{B}} = \mathrm{var}(\delta_j)$, we can introduce constraints analogous to those in (3.5), so that an expression more accurate than (3.7) is

$$y_{ij} = \mathbf{x}_j^{(1)} \beta^{(1)} + \mathbf{x}_j^{(2)} \beta^{(2)} + \mathbf{z}_j \delta_j + \varepsilon_{ij}, \tag{3.8}$$

where $\mathbf{z}$ is a subset of the variables in $\mathbf{x}^{(1)}$. The interpretation in terms of varying regression slopes also carries over to the longitudinal setting. The within-subject

regression slopes with respect to the variables in $\mathbf{z}$ vary, and with respect to their complement in $\mathbf{x}^{(1)}$ are constant.

The vector $\mathbf{x}^{(1)}$ contains the variable(s) that represent time. For linear within-subject regressions, time is represented by a single variable, but growth may follow any other pattern. To cater for the possibilities, transformations of the time have to be included in $\mathbf{X}^{(1)}$, and some of them also in $\mathbf{Z}$. Invariance with respect to linear transformations dictates that a variable included in $\mathbf{Z}$ should be included also in $\mathbf{X}^{(1)}$. Further, when a hierarchy is defined for the variables in $\mathbf{X}$, such as in polynomial regression, then this hierarchy should also be reflected in the model choice. For example, if the cubic term, t^3, is included in $\mathbf{X}^{(1)}$, then so should be the linear and quadratic terms. Similarly, if t^3 is included in $\mathbf{Z}$, then so should be the linear and quadratic terms. However, if t^3 is included in $\mathbf{X}^{(1)}$, it does not have to be included in $\mathbf{Z}$ although if t^2 is included, then so should be the linear term t.

The two-level model (Longford, 1993; Verbeke and Molenberghs, 2000; and Goldstein, 2003) can be applied when observations are made at a given (fixed) set of time points, but some limitations arise for the combination $\mathbf{z}_j\delta_j$. For r time points, the largest possible dimension of δ_j is r. The multivariate model in (3.6) corresponds to r-variate δ_j with $\mathbf{Z}$ comprising the unity (intercept) and the indicators of the categories 2, 3, ..., r. Other options correspond to a reparametrisation of such a vector $\mathbf{Z}$. When the observations are not made in a regimented fashion, the number of variables in $\mathbf{Z}$ may still have to be restricted. To see this, consider a design with time points that for every subject are drawn from the same set, such as 1, ..., 10, but not every subject has all the ten observations. Then $\mathbf{Z}$ should not contain more than ten covariates (columns). A direct analogy can be drawn with the models for the analysis of covariance (ANCOVA). The models in (3.8) differ from them solely by associating the subject specific deviations δ_j with randomness; in ANCOVA they are (fixed) parameters, subject only to the constraints of identifiability.

A subject-level variable $X^{(2)}$ is by definition constant within subjects, and so the within-subject regression with respect to $X^{(2)}$ is not well defined. The only reason why such a variable might be included in $\mathbf{Z}$ is to model variance heterogeneity — the dependence of the variance on the covariates. In general, for the model in (3.8), we have the identities

$$\begin{aligned} \mathrm{var}(y_{ij}) &= \sigma^2 + \mathbf{z}_{ij}\Sigma_{\mathrm{B}}\mathbf{z}_{ij}^{\top} \\ \mathrm{cov}(y_{ij}, y_{i'j}) &= \mathbf{z}_{ij}\Sigma_{\mathrm{B}}\mathbf{z}_{i'j}^{\top} \end{aligned} \tag{3.9}$$

for $i \neq i'$. Both expressions are quadratic functions of the components of $\mathbf{z}$. Therefore, exploring the properties of $\mathrm{var}(y)$ and $\mathrm{cov}(y_1, y_2)$ as functions of $\mathbf{z}$ is relatively simple, although the components of $\mathbf{z}$ may be interrelated, such as the indicator variables for a categorical variable, or the linear and quadratic terms of a polynomial. The range of the values of the time t is usually limited, so we can evaluate $\mathrm{var}(y)$ as a function of t unambiguously when $\mathbf{z}$ contains only functions of time. Otherwise we have to consider a few (typical) values of the other variables and evaluate $\mathrm{var}(y)$ for each of them.

3.5.1 Example

Figure 3.1 gives an example of a longitudinal dataset with observations at time points 1, 2, ..., 12 for 40 subjects. The outcomes are generated according to the model in (3.8) with no covariates, except for the time and its transformations. For the regression $\mathbf{x}^{(1)}\beta$ we use a cubic polynomial in t, and for the variation $\mathbf{z}\delta_j$ a quadratic polynomial in t. Thus, the values $\mathbf{x}^{(1)}\beta + \mathbf{z}\delta_j$, $j = 1,\ldots,40$, are cubic polynomials with the same cubic coefficient but different quadratic (and linear and absolute) coefficients. The data are generated with

$$\beta = (1,\ 0.3,\ 0.024,\ 0.0011)^\top,$$

$\sigma^2 = 0.25$ and

$$\Sigma = \begin{pmatrix} 0.60000 & 0.04000 & 0.00030 \\ 0.04000 & 0.02000 & 0.00015 \\ 0.00030 & 0.00015 & 0.00009 \end{pmatrix}.$$

The curves (trajectories) $\mathbf{x}^{(1)}\beta + \mathbf{z}\delta_j$ are plotted in panel A. We refer to them as smooth or underlying trajectories, because they are devoid of the inexplicable contribution ε. This random term is commonly referred to as an error. In most contexts, this label is inappropriate and misleading. It would be appropriate if the model we specify were correct (as it is in a simulation) and if all subjects behaved according to this model with σ^2, and the elementary-level deviations ε arose as a result of an imperfect measurement process. In most cases, the model is incorrect, and a particular positive value of σ^2 is appropriate because subjects do not behave according to any conceivable formula, but there are some equations (models) that approximate the behaviour reasonably well. The approximation is in error, not the behaviour.

Panel B presents the trajectories as they would be observed, made coarse by the elementary-level deviations ε. It is difficult to infer the patterns of the trajectories, smooth or coarse, from the parameter values in β, Σ and σ^2, except perhaps for the extent of the average curvature (from β_3) and the extent of inexplicability (from σ^2); using a simulated sample from the fitted model is much more reliable. Such a sample, replicated several times, also has an important diagnostic value, as discussed in Section 3.5.3.

The variance of an observation, as a function of time t,

$$\mathrm{var}(y\,|\,t) = \sigma^2 + (1,\ t,\ t^2)\,\Sigma\,(1,\ t,\ t^2)^\top,$$

is drawn in panel C, together with the indication of σ^2 as its 'constant' contributor (drawn by dashes). There is no profound reason why the elementary-level variance should be constant; it is merely a convenient assumption. Without it, we would have to posit a particular form for how σ^2 depends on t. Alternatives plausible in some settings are that the correlation of two outcomes of the same subject is constant, or the ratio of the within- and between-subject variances is constant.

There is a trade-off between a within-subject variance σ_t^2 and the subject-level matrix Σ. That is, up to a point, a change in one or several values of σ_t^2 can be

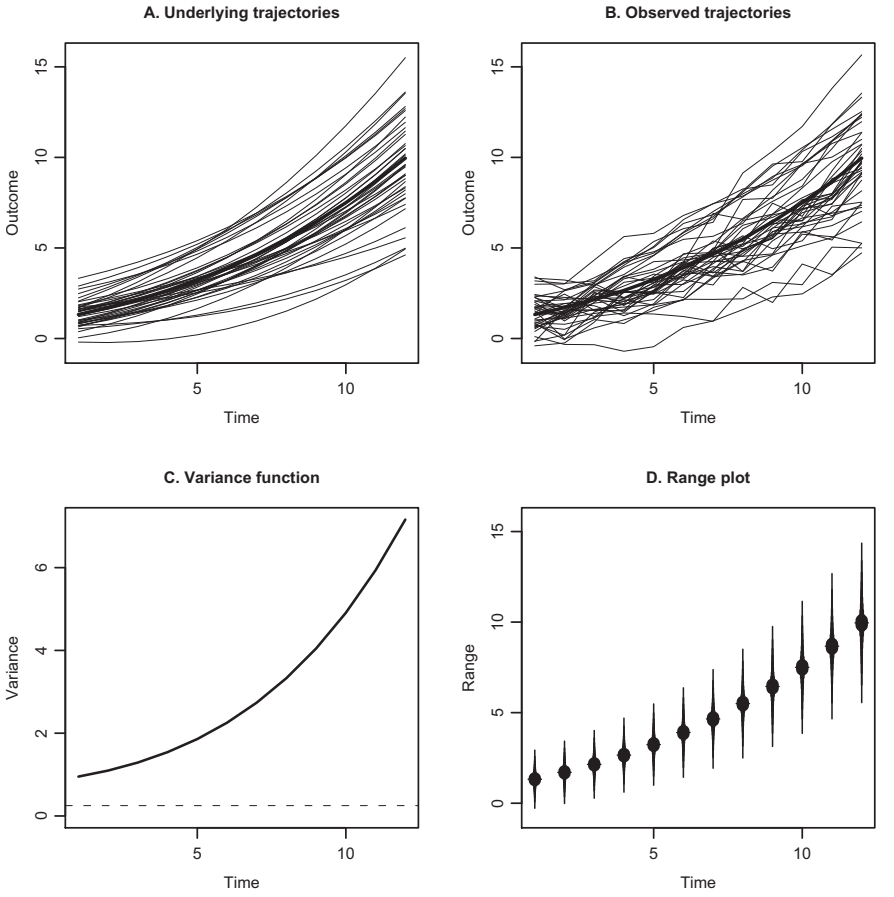

Fig. 3.1 Graphical representation of balanced longitudinal data. A simulated example.

compensated by changes in Σ, so long as the 'new' Σ remains non-negative definite. In principle, σ^2 is identifiable from the data, because it represents the *independent* contribution to the variance $\mathrm{var}(y)$. The variance matrix Σ characterises the covariance structure of the observations of a subject. However, large samples are required to separate the two components of variance, σ_t^2 and Σ, with any meaningful reliability.

Panel D of Figure 3.1 illustrates the distributions of the outcomes within the time points. It does not contain all the information about the underlying process, because it gives no indication of the covariance structure of the outcomes. In this respect, there is no replacement for the simulated trajectories in panels A and B.

Nonlinear transformations can alter the pattern of the trajectories substantially, from convex to approximately linear to concave. With a nonlinear transformation, we manipulate the underlying distributions (normal for ε and δ), the covariance

structure, and the pattern of the variances σ_t^2. In theory, only one family (class of equivalence) of transformations yields normally distributed outcomes, so arranging for all the distributional assumptions to hold is next to impossible. In practice, there is a considerable leeway in the choice of a transformation to make the assumptions of normality palatable. In fact, in many settings we can focus on transformations that bring about variance homogeneity (independence of the variances var(y) and σ_t^2 on time t) as well.

3.5.2 The Time-Selection Process

In many longitudinal studies, the values of the time points t are not set by design, prior to data collection. For example, a study may rely on subscribing individuals turning up at a given location for a particular service, such as health care, advice with jobs search, a form of entertainment, and similar. In such settings, the realised values of t may be *informative*, and the process that generates its values *nonignorable*. The observed data are not a good reflection of the process we set out to study.

This problem does not have a solution, in that there is no straightforward way of adjusting the analysis so that it would be suitable for inferences about the entire evolution of the outcome variables, or about the values of the outcome variables at time points selected by design, with the subjects exercising no choice in the matter.

3.5.3 Simulation-Based Diagnostics

Established methods for model diagnostics are difficult to adapt for longitudinal analysis because of a combination of concerns about normality, appropriate covariance structure and heteroscedasticity. The following generic procedure, introduced by Rubin (1984), can be applied. We define a data summary called *feature*; this can be a single quantity, a vector, a table, a diagram, or their combination (a *multifeature*). We evaluate (or apply) this feature to the realised dataset, thus obtaining the *realised* feature. Next, we simulate datasets from the model fit using the same design (sample sizes and values of the covariates) as the realised data, and evaluate the feature on each replicate dataset. We shuffle the one realised and the several simulated features, and ask a third party (a colleague) to identify one of them as being exceptional. If he or she points to the realised feature (without knowing that it is based on the real dataset and the others are not), we conclude that the model is not appropriate, because if it were, as it is with the simulated data, then the features would not look (or be) different. It is advantageous to generate 19, 49 or 99 replicate datasets, so that we would have 20, 50 or 100 datasets and could relate the probability of identifying the realised dataset by chance to the size of a test in hypothesis testing. The price for greater accuracy is having to generate a greater number of

replicates (a serious problem only with very large datasets), and presenting a more cumbersome task for the colleague. See Longford (2001) for an example.

3.6 Multivariate Observations at Time Points

Suppose the observations of a set of variables at a time point t are well described by a multivariate normal distribution $\mathcal{N}(\mu_t, \Sigma_t)$, specific to the time point t. We are concerned about the evolution of these distributions across the time points. This entails specifying models for the vectors of expectations μ_t and variance matrices Σ_t, but also for the correlation structure of vectors of observations at distinct (consecutive) time points. This is necessary even in the stationary case, when the matrices Σ_t are identical. For example, the assumption that vectors of outcomes $\mathbf{y}_t$ and $\mathbf{y}_{t'}$ are independent for distinct time points $t \neq t'$ is in most settings untenable, and so is the assumption of perfect correlation, $\mathbf{C}_{tt'} = \mathrm{cov}(\mathbf{y}_t, \mathbf{y}_{t'}) = \Sigma_t$.

Multivariate longitudinal outcomes are represented by a matrix of variables $\mathbf{Y}$, comprising vectors of variables $\mathbf{y}_t$ at a time point as its rows and the time series of univariate longitudinal outcomes $\mathbf{y}^{(k)} = \left(y_1^{(k)}, y_2^{(k)}, \ldots y_T^{(k)}\right)^\top$ as its columns. An ideal solution for the correlation structure across the time points would allow an (arbitrary) univariate longitudinal model for each component $\mathbf{y}^{(k)}$ and a rich variety of dependence structures implied by the covariance matrices $\mathbf{C}_{tt'}$. Of course, imposing constraints such as non-negative covariances in $\mathbf{C}_{tt'}$ and higher correlations for pairs of time points t and t' in greater proximity, is reasonable in most contexts. We seek models mainly for short time series (small T), so we are not concerned about stationarity and other properties that are related to large T.

3.6.1 Autoregression

The univariate autoregression has an obvious multivariate analogue,

$$\mathbf{y}_{t+1} = \mathbf{a}_t + \mathbf{B}_t \mathbf{y}_t + \varepsilon_t, \tag{3.10}$$

where $\mathbf{a}_t$ is a vector and $\mathbf{B}_t$ a matrix of coefficients and ε_t a centred random vector independent of $\mathbf{y}_1, \ldots, \mathbf{y}_t$. To maintain multivariate normality, we assume that $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \Xi)$ and $\mathbf{y}_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$. Then

$$\begin{aligned} \mathbf{C}_{t,t} &= \Sigma_t = \mathbf{B}_t \Sigma_{t-1} \mathbf{B}_t^\top + \Xi \\ \mathbf{C}_{t,t+1} &= \Sigma_t \mathbf{B}_t^\top. \end{aligned}$$

An important special case arises when $\mathbf{B}_t$ is diagonal. This does not correspond to independent autoregressions, because dependence is still injected by the covariance

structure of ε_t, as well as the initial covariance matrix Σ_1. When $\mathbf{y}_t$ comprises closely related variables, the components of ε_t are correlated.

3.6.2 Moving Average

The univariate moving average model has a similar extension for multivariate outcomes. Each time point t is associated with an independent random vector ε_t with centred multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Xi)$, and the vector of outcomes is assumed to be generated according to the model

$$\mathbf{y}_t = \mu_t + \mathbf{A}_0\varepsilon_t + \mathbf{A}_1\varepsilon_{t-1} \tag{3.11}$$

for some matrices of constants $\mathbf{A}_0$ and $\mathbf{A}_1$. A joint distribution has to be specified for the start of the series, $(\mathbf{y}_1, \mathbf{y}_2)$. The essential multivariateness of such a moving average arises as a result of the covariance structure of Ξ *combined* with the non-zero off-diagonal elements of $\mathbf{A}_0$ and $\mathbf{A}_1$. The models in (3.10) and (3.11) are for the respective first-order autoregressive and moving-average models. Their generalisation to higher-order models is straightforward. However, such models are of limited use with short time series typically encountered in longitudinal analysis. Autoregression and moving average yield distinct sets of models, so that, at least in principle, the issue of distinguishing between them, e.g., by hypothesis testing or information criteria, may arise. In practice, such tests have limited power even in the univariate case, so the data-based choice between them is unlikely to be feasible in a multivariate setting. The two kinds of models can be combined, in analogy with the univariate case.

3.6.3 Two-Level Models

The multivariate version of the compound symmetry model in (3.3) is

$$\mathbf{Y}_j = \mathbf{X}_j\mathbf{B} + \mathbf{1}\delta_j^\top + \mathbf{E}_j,$$

where $\mathbf{Y}_j$ is the $T \times K$ (times-by-variables) matrix of outcomes for subject j, $\mathbf{X}_j$ the corresponding matrix of covariates, $\mathbf{B}$ is a matrix of regression parameters, δ_j a random sample from a multivariate normal distribution (one vector per subject), $\mathbf{t}_j$ is the vector of time points, and $\mathbf{E}_j$ a matrix; the rows $\mathbf{e}_{ij}$ of $\mathbf{E}_j$ are mutually independent random vectors (a multivariate random sample), both within a matrix $\mathbf{E}_j$ and across them, from another multivariate normal distribution. The two random samples, δ_j (for subjects) and $\mathbf{e}_{ij}$ (for occasions within subjects), are mutually independent.

A column of the matrix $\mathbf{X}_j$ is time and some others are its transformations. These can be associated with subject-level variation by the model

$$\mathbf{Y}_j = \mathbf{X}_j\mathbf{B} + \mathbf{1}\delta_j^{(0)\top} + \mathbf{t}_j\delta_j^{(1)\top} + \mathbf{E}_j, \tag{3.12}$$

where the matrix $\Delta_j = (\delta_j^{(0)}, \delta_j^{(1)})$ is a random sample from a multivariate normal distribution. In this model, the subjects have different associations with time (varying coefficients $\delta_j^{(1)}$). The model can be supplemented with transformations of time, injecting more flexibility in how the values of the variables within subjects evolve. Covariances in $\mathrm{var}(\Delta_j)$ are essential, because the evolution of the variables is unlikely to be independent (unrelated). Of course, $\mathbf{E}_j$ induces some dependence among the rows of $\mathbf{Y}_j$, but we can regard $\mathrm{E}(\mathbf{Y}_j\,|\,\mathbf{E}_j) = \mathbf{X}_j\mathbf{B} + \mathbf{1}\delta_j^{(0)\top} + \mathbf{t}_j\delta_j^{(1)\top}$ as an underlying trend, and study its dependence structure.

We distinguish among variables defined for subjects, which are represented in each $\mathbf{X}_j$ by a column of constants, and variables defined for occasions. The time, represented in $\mathbf{X}_j$ by a column $\mathbf{t}_j$, is one such variable. Variables that are not functions of time, but are recorded on every occasion (observation), may also be included in $\mathbf{X}_j$. Such variables are usually called *time-varying*. They can be associated with subject-level variation to model the varying within-subject regressions of the outcomes on them. Associations of variables with the outcomes have to be interpreted with care when the values of these variables are recorded passively, without (experimental) control over them. In the framework of causal analysis, they may be 'intermediate' variables, affected by the earlier outcomes, and so their associations with the outcomes differ from the causal effects of these variables.

3.7 Maximum Likelihood Estimation

Maximisation of the likelihood with the normality assumptions is conceptually simple and is relatively easy to implement because the likelihood for all the models we consider has an analytical form. Some difficulties are caused by the large number of parameters some of which are connected by the assumed structures. The constraints of nonnegative definiteness are difficult to enforce. Other difficulties arise in the model specification, because there is no obvious way of defining a sequence of nested models that would represent gradual increase in model complexity. Connection of the substantive information with such constraints is particularly difficult to establish. In principle, we could define the joint distribution of all the outcomes directly. In such a definition, it is difficult to reflect the structure of observations within time points.

Likelihood maximisation involves iterative procedures, and these require a (good) initial solution. Initial solutions are frequently the fits of some very simple submodels which are obtained by a simple algorithm. A practical initial solution for fitting the model in (3.12) or its generalisations is the set of univariate multilevel model fits. These themselves require iterations, but they are much simpler than a 'multivariate' iteration. The univariate model fits are useful also for exploring informally the choice of models for the marginals, the components of $\mathbf{X}_j\mathbf{B}$.

Large variance matrices (of model parameters) are estimated by an algorithm that does not internally respects the nonnegative definiteness of the estimated variance matrices. In a large (estimated) matrix $\hat{\Omega}$, the presence of a negative eigenvalue is not obvious, so the problem might be ignored, until we come across a negative value of a quadratic form $\mathbf{c}^\top \hat{\Omega} \mathbf{c}$ for a vector of constants $\mathbf{c}$ or wish to draw a random sample from the fitted distribution. The constraints of nonnegative definiteness are difficult to implement in a full-proof fashion, because they involve a trade-off between slowing down the convergence rate and ensuring that the solution moves smoothly from one iteration to the next along (or close to) the boundary of the parameter space defined by nonnegative definiteness.

Alternative solutions estimate decompositions of the variance matrices, such as the Cholesky or single-value, but the structures we want to impose on the variance matrices are very difficult to convert to the constraints on these decompositions.

There is no comprehensive software for multivariate random coefficient models, but software for univariate models can be adapted for the purpose. `MLwin` (Rasbash *et al*, 2005) and the software `nlme` described in Pinheiro and Bates (2000) are well suited for this purpose. For methods, examples and general background, we recommend Diggle *et al* (2002). Laird and Ware (1982) is a paper of historical importance, outlining the application of random coefficient models for longitudinal analysis. There is extensive Bayesian literature on longitudinal analysis, much of it centred around or using the `WinBugs` software (`wwww.mrc-bsu.ca.ac.uk/bugs`).

3.7.1 Graphics – Initial Data Exploration

The first step in an initial exploration of the data is to plot the trajectories (evolutions) for each variable separately. The next step entails representing the dependence of the observations across the variables. Plotting the trajectories of the distinct variables side-by-side, with the subject marked for each trajectory is effective only for a few subjects (e.g., a random sample drawn from the data), so that the trajectories of a subject could be easily identified in the adjacent panels. In multivariate models with random slopes, the variances and correlations of the observations are time-specific, and so we can study their evolution by plotting them as functions of time. This can be effectively implemented by a matrix plot (function `pairs` in `R`), with the variances plotted in the diagonal panels and the correlations plotted in the off-diagonal panels. More information is displayed when the correlations are plotted under the diagonal and the covariances above it.

Figure 3.2 presents a bivariate longitudinal dataset. The relatively smooth lines in the top panels are for the underlying trends, devoid of the within-subject variation. The average trend (the regressions) are drawn by thick lines in the top panels. They enable, however crudely, to gain an impression of the correlation of the two outcomes (components). Comparisons within columns help us to assess the impact of the within-subject variation, commonly interpreted as noise or error, although an attribution of ε to a replication-specific random variable (due to the subject's

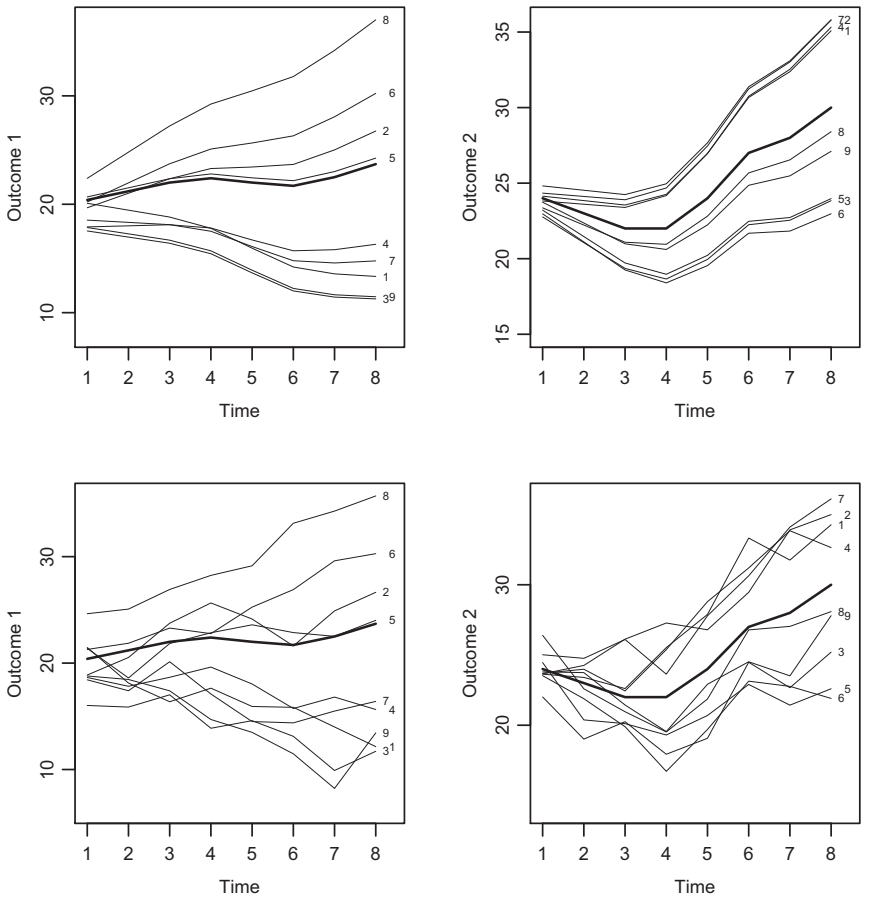

Fig. 3.2 A bivariate longitudinal dataset. A simulated example. The top panels display the underlying trends and the bottom panels the values of the observations. The thick solid line indicates the marginal (population) mean.

inconsistency in the response or imperfection of the measurement/recording process) is not always warranted.

In this example, the subject-level variance matrix was specified as

$$\Omega = \begin{pmatrix} 1.042 & -0.112 & 0.376 & -0.028 \\ -0.112 & 1.602 & 0.104 & 0.268 \\ 0.376 & 0.104 & 0.928 & 0.026 \\ -0.028 & 0.268 & 0.026 & 0.337 \end{pmatrix},$$

constructed from an eigenvalue decomposition to ensure nonnegative definiteness. The additional space in the display separates the rows and columns that correspond

to the outcomes, and within each 2×2 matrix, the first component corresponds to the intercept and the second to (linear) time. The within-subject variance matrix is

$$\Sigma = \begin{pmatrix} 1.8 & 1.0 \\ 1.0 & 1.4 \end{pmatrix},$$

and the vectors of the population means for the two components are

$$\mu_1 = (20.4, 21.2, 22.0, 22.4, 22.0, 21.7, 22.5, 23.7)^\top$$
$$\mu_2 = (24.0, 23.0, 22.0, 22.0, 24.0, 27.0, 28.0, 30.0)^\top.$$

Figure 3.3 summarizes the marginal distributions graphically, highlighting the increasing variation with time.

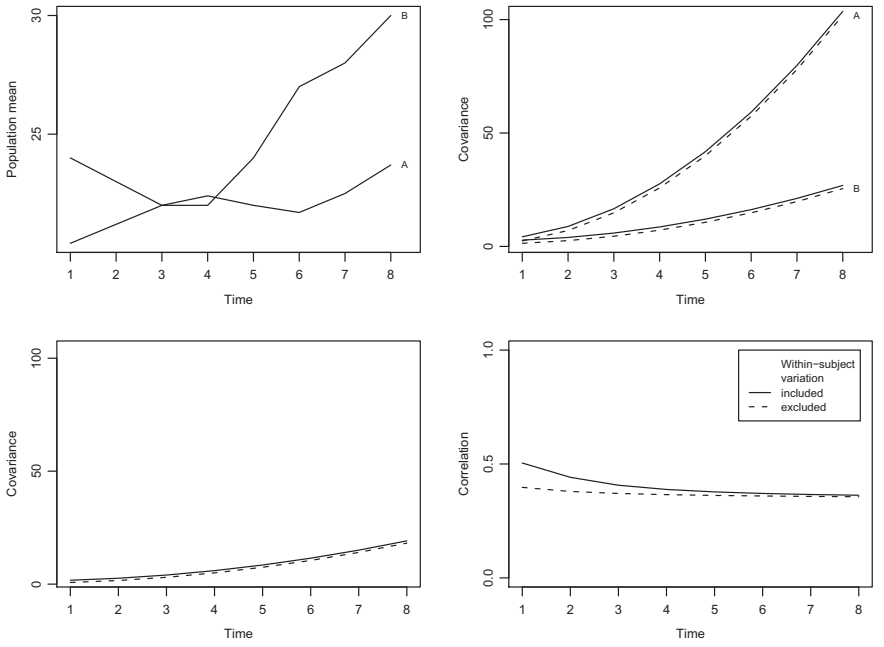

Fig. 3.3 The (marginal) summaries of a bivariate longitudinal series: trend (expectations), variances, covariances and correlations. Simulated data, with the parameters given in the text.

3.8 Discussion

Longitudinal analysis refers to analysis that involves a time dimension. In this respect it is multivariate, although it may involve other aspects of multivariateness, such as several outcomes being observed at each time point. Longitudinal data comprise repeated observations on subjects, so that their change (growth, decay or development) can be studied. The temporal dependence can be accounted for by regression or correlation structures, or their combinations, and the subject-to-subject variation by random coefficients. In the model construction, for estimation and prediction, we can draw on models for time series (autoregression and moving average) and for random coefficients. These are most conveniently specified with the assumptions of normality and linearity, for which estimation procedures are relatively simple, based on maximum likelihood. Transformations and the generalized linear modelling framework cater for departures from normality.

Designing longitudinal studies and dealing with nonresponse, and designing studies which anticipate nonresponse, are challenging problems that do not have a universal solution because of the intricate interplay of the correlation structure of the outcome variables with the quality of the estimation. Survey expenses are an important consideration, especially in studies that take place over a long period of time (several years) and in populations that, in general, do not have a stake in the survey and regard responding as a distraction from their everyday affairs. Methods for dealing with nonresponse and with data that do not fit into neat rectangular data structures have an important role in the analysis of such surveys.

References

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Ser. A, 158,* 419-466.

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. L. (2002). *Analysis of longitudinal data* (2nd ed.). Oxford: Oxford University Press.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Ser. B, 57,* 45-97.

Francis, B., Green, M., & Payne, C. (1993). *The GLIM system. Release 4 manual.* Oxford: Oxford University Press.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: E. Arnold.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945-970.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38,* 963-974.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Longford, N. T. (1993). *Random coefficient models.* Oxford: Oxford University Press.

Longford, N. T. (2001). Simulation-based diagnostics in random-coefficient models. *Journal of the Royal Statistical Society, Ser. A, 64,* 259-273.

Longford, N. T. (2007). *Studying human populations. An advanced course in statistics.* New York: Springer-Verlag.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in* S *and* Splus. New York: Springer-Verlag.

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2005). MLwin version 2.02. Centre for Multilevel Modelling, University of Bristol.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics, 12,* 1151-1172.

Rubin, D. B. (2005). Causal inference using potential outcomes: design, modelling, decisions. 2004 Fisher Lecture. *Journal of the American Statistical Association, 100,* 32-331.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York: Springer-Verlag.