Statistics for Biology and Health

David G. Kleinbaum · Mitchel Klein

Logistic Regression

A Self-Learning Text

Third Edition

Springer

Statistics for Biology and Health

For other titles published in this series, go to
http://www.springer.com/series/2848

David G. Kleinbaum   Mitchel Klein

Logistic Regression

A Self-Learning Text

Third Edition

With Contributions by
Erica Rihl Pryor

Springer

David G. Kleinbaum
Mitchel Klein
Department of Epidemiology
Emory University
Rollins School of Public Health
Atlanta, GA 30322
USA
dkleinb@sph.emory.edu
mklein@sph.emory.edu

ISBN: 978-1-4419-1741-6 e-ISBN: 978-1-4419-1742-3
DOI 10.1007/978-1-4419-1742-3
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2009943538

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To

Edna Kleinbaum

and

Rebecca Klein

Contents

Preface

This is the third edition of this text on logistic regression methods, originally published in 1994, with its second edition published in 2002.

As in the first two editions, each chapter contains a presentation of its topic in "lecture-book" format together with objectives, an outline, key formulae, practice exercises, and a test. The "lecture book" has a sequence of illustrations, formulae, or summary statements in the left column of each page and a script (i.e., text) in the right column. This format allows you to read the script in conjunction with the illustrations and formulae that highlight the main points, formulae, or examples being presented.

This third edition has expanded the second edition by adding three new chapters and a modified computer appendix. We have also expanded our overview of modeling strategy guidelines in Chap. 6 to consider causal diagrams. The three new chapters are as follows:

Chapter 8: Additional Modeling Strategy Issues
Chapter 9: Assessing Goodness of Fit for Logistic Regression
Chapter 10: Assessing Discriminatory Performance of a Binary Logistic Model: ROC Curves

In adding these three chapters, we have moved Chaps. 8 through 13 from the second edition to follow the new chapters, so that these previous chapters have been renumbered as Chaps. 11–16 in this third edition. To clarify this further, we list below the previous chapter titles and their corresponding numbers in the second and third editions:

Chapter Title	*Chapter #* *2nd Edition*	*Chapter #* *3rd Edition*
Analysis of Matched Data Using Logistic Regression	8	11
Polytomous Logistic Regression	9	12
Ordinal Logistic Regression	10	13
Logistic Regression for Correlated Data: GEE	11	14
GEE Examples	12	15
Other Approaches for Analysis of Correlated Data	13	16

New Chap. 8 addresses five issues on modeling strategy not covered in the previous two chapters (6 and 7) on this topic:

Issue 1: Modeling Strategy When There Are Two or More Exposure Variables
Issue 2: Screening Variables When Modeling
Issue 3: Collinearity Diagnostics
Issue 4: Multiple Testing
Issue 5: Influential Observations

New Chap. 9 addresses methods for assessing the extent to which a binary logistic model estimated from a dataset predicts the observed outcomes in the dataset, with particular focus on the deviance statistic and the Hosmer-Lemeshow statistic.

New Chap. 10 addresses methods for assessing the extent that a fitted binary logistic model can be used to distinguish the observed cases from the observed noncases, with particular focus on ROC curves.

The modified appendix, Computer Programs for Logistic Regression, updates the corresponding appendix from the second edition. This appendix provides computer code and examples of computer programs for the different types of logistic models described in this third edition. The appendix is intended to describe the similarities and differences among some of the most widely used computer packages. The software packages considered are SAS version 9.2, SPSS version 16.0, and Stata version 10.0

Suggestions for Use

This text was originally intended for self-study, but in the 16 years since the first edition was published, it has also been effectively used as a text in a standard lecture-type classroom format. Alternatively, the text may be used to supplement material covered in a course or to review previously learned material in a self-instructional or distance-learning format. A more individualized learning program may be particularly suitable to a working professional who does not have the time to participate in a regularly scheduled course.

The order of the chapters represents what the authors consider to be the logical order for learning about logistic regression. However, persons with some knowledge of the subject can choose whichever chapter appears appropriate to their learning needs in whatever sequence desired.

The last three chapters (now 14–16) on methods for analyzing correlated data are somewhat more mathematically challenging than the earlier chapters, but have been written to logically follow the preceding material and to highlight the principal features of the methods described rather than to give a detailed mathematical formulation.

In working with any chapter, the user is encouraged first to read the abbreviated outline and the objectives, and then work through the presentation. After finishing the presentation, the user is encouraged to read the detailed outline for a summary of the presentation, review key formulae and other important information, work through the practice exercises, and, finally, complete the test to check what has been learned.

Recommended Preparation

The ideal preparation for this text is a course on quantitative methods in epidemiology and a course in applied multiple regression. The following are recommended references on these subjects with suggested chapter readings:

Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H., *Epidemiologic Research: Principles and Quantitative Methods*, Wiley, New York, 1982, Chaps. 1–19.

Kleinbaum, D.G., Kupper, L.L., Nizam, A., and Muller, K.A., *Applied Regression Analysis and Other Multivariable Methods, Fourth Edition*, Duxbury Press/Cengage Learning, Pacific Grove, 2008, Chaps. 1–16.

Kleinbaum, D.G., *ActivEpi- A CD-Rom Text*, Springer, New York, 2003, Chaps. 3–15.

A first course on the principles of epidemiologic research would be helpful since this text is written from the perspective of epidemiologic research. In particular, the learner should be familiar with basic characteristics of epidemiologic study designs and should have some understanding of the frequently encountered problem of controlling/adjusting for variables.

As for mathematics prerequisites, the learner should be familiar with natural logarithms and their relationship to exponentials (powers of *e*) and, more generally, should be able to read mathematical notation and formulae.

Atlanta, GA

David G. Kleinbaum
Mitchel Klein

Acknowledgments

David Kleinbaum and Mitch Klein continue to thank Erica Pryor at the School of Nursing, University of Alabama-Birmingham, for her many important contributions, including editing, proofing, and computer analyses, to the second edition. We also want to thank Winn Cashion for carefully reviewing the three new chapters of the third edition.

We also thank our wives, Edna Kleinbaum and Becky Klein, for their love, friendship, advice, and support as we were writing this third edition. In appreciation, we are dedicating this edition to both of them.

Atlanta, GA

David G. Kleinbaum
Mitchel Klein

1 Introduction to Logistic Regression

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_1, © Springer Science+Business Media, LLC 2010

Introduction

This introduction to logistic regression describes the reasons for the popularity of the logistic model, the model form, how the model may be applied, and several of its key features, particularly how an odds ratio can be derived and computed for this model.

As preparation for this chapter, the reader should have some familiarity with the concept of a mathematical model, particularly a multiple-regression-type model involving independent variables and a dependent variable. Although knowledge of basic concepts of statistical inference is not required, the learner should be familiar with the distinction between population and sample, and the concept of a parameter and its estimate.

Abbreviated Outline

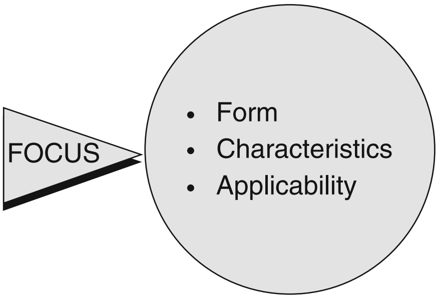
The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

Objectives

Upon completing this chapter, the learner should be able to:

1. Recognize the multivariable problem addressed by logistic regression in terms of the types of variables considered.
2. Identify properties of the logistic function that explain its popularity.
3. State the general formula for the logistic model and apply it to specific study situations.
4. Compute the estimated risk of disease development for a specified set of independent variables from a fitted logistic model.
5. Compute and interpret a risk ratio or odds ratio estimate from a fitted logistic model.
6. Identify the extent to which the logistic model is applicable to followup, case-control, and/or cross-sectional studies.
7. Identify the conditions required for estimating a risk ratio using a logistic model.
8. Identify the formula for the logit function and apply this formula to specific study situations.
9. Describe how the logit function is interpretable in terms of an "odds."
10. Interpret the parameters of the logistic model in terms of log odds.
11. Recognize that to obtain an odds ratio from a logistic model, you must specify **X** for two groups being compared.
12. Identify two formulae for the odds ratio obtained from a logistic model.
13. State the formula for the odds ratio in the special case of (0, 1) variables in a logistic model.
14. Describe how the odds ratio for (0, 1) variables is an "adjusted" odds ratio.
15. Compute the odds ratio, given an example involving a logistic model with (0, 1) variables and estimated parameters.
16. State a limitation regarding the types of variables in the model for use of the odds ratio formula for (0, 1) variables.

Presentation



This presentation focuses on the basic features of logistic regression, a popular mathematical modeling procedure used in the analysis of epidemiologic data. We describe the *form* and key *characteristics* of the model. Also, we demonstrate the *applicability* of logistic modeling in epidemiologic research.

I. The Multivariable Problem

$$E \xrightarrow{?} D$$

We begin by describing the multivariable problem frequently encountered in epidemiologic research. A typical question of researchers is: What is the relationship of one or more exposure (or study) variables (E) to a disease or illness outcome (D)?

EXAMPLE

$D_{(0,\,1)} = \text{CHD}$

$E_{(0,\,1)} = \text{SMK}$

$$\text{SMK} \longrightarrow \text{CHD}$$

"control for"

$C_1 = \text{AGE}$

$C_2 = \text{RACE}$

$C_3 = \text{SEX}$

To illustrate, we will consider a dichotomous disease outcome with 0 representing *not diseased* and 1 representing *diseased*. The dichotomous disease outcome might be, for example, coronary heart disease (CHD) status, with subjects being classified as either 0 ("without CHD") or 1 ("with CHD").

Suppose, further, that we are interested in a single dichotomous exposure variable, for instance, smoking status, classified as "yes" or "no". The research question for this example is, therefore, to evaluate the extent to which smoking is associated with CHD status.

To evaluate the extent to which an exposure, like smoking, is associated with a disease, like CHD, we must often account or "control for" additional variables, such as age, race, and/or sex, which are not of primary interest. We have labeled these three control variables as C_1, C_2, and C_3.

$$\underbrace{E,\ C_1,\ C_2,\ C_3}_{\text{independent}} \xrightarrow{?} \underset{\text{dependent}}{D}$$

In this example, the variable E (the exposure variable), together with C_1, C_2, and C_3 (the control variables), represents a collection of *independent* variables that we wish to use to describe or predict the *dependent* variable D.

Independent variables:
$X_1, X_2, \ldots, X_k$

More generally, the independent variables can be denoted as X_1, X_2, and so on up to X_k, where k is the number of variables being considered.

Xs may be Es, Cs, or combinations

We have a *flexible* choice for the Xs, which can represent any collection of exposure variables, control variables, or even combinations of such variables of interest.

EXAMPLE	
$X_1 = E$	$X_4 = E \times C_1$
$X_2 = C_1$	$X_5 = C_1 \times C_2$
$X_2 = C_2$	$X_6 = E^2$

For example, we may have the following:

X_1 equal to an exposure variable E

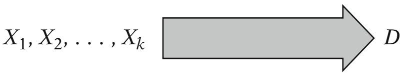
X_2 and X_3 equal to control variables C_1 and C_2, respectively

X_4 equal to the product $E \times C_1$

X_5 equal to the product $C_1 \times C_2$

X_6 equal to E^2

The Multivariable Problem



Whenever we wish to relate a set of Xs to a dependent variable, like D, we are considering a *multivariable problem*. In the analysis of such a problem, some kind of *mathematical model* is typically used to deal with the complex interrelationships among many variables.

The analysis:
mathematical model

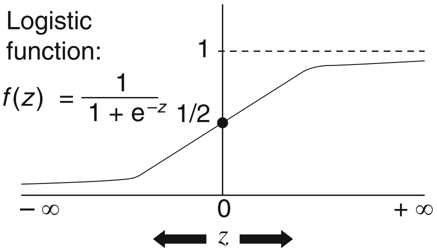
Logistic model:
dichotomous D

Logistic regression is a mathematical modeling approach that can be used to describe the relationship of several Xs to a *dichotomous* dependent variable, such as D.

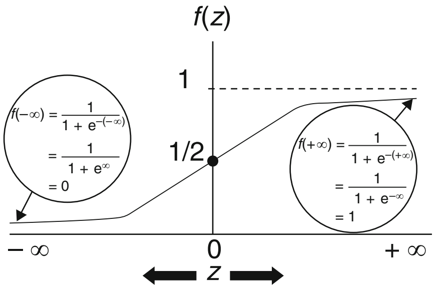
Logistic is most popular

Other modeling approaches are possible also, but logistic regression is by far the most *popular* modeling procedure used to analyze epidemiologic data when the illness measure is dichotomous. We will show why this is true.

II. Why Is Logistic Regression Popular?



To explain the popularity of logistic regression, we show here the *logistic function*, which describes the mathematical form on which the *logistic model* is based. This function, called $f(z)$, is given by 1 over 1 plus e to the minus z. We have plotted the values of this function as z varies from $-\infty$ to $+\infty$.



Notice, in the balloon on the left side of the graph, that when z is $-\infty$, the logistic function $f(z)$ equals 0.

On the right side, when z is $+\infty$, then $f(z)$ equals 1.

Range: $0 \leq f(z) \leq 1$

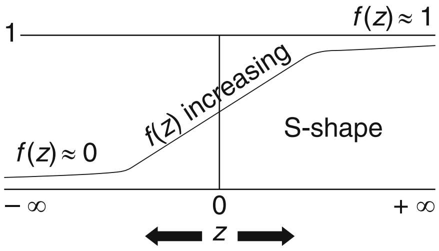
Thus, as the graph describes, the *range* of $f(z)$ is between 0 and 1, regardless of the value of z.

$0 \leq$ probability ≤ 1 (individual risk)

The fact that the logistic function $f(z)$ *ranges between* 0 *and* 1 is the primary reason the logistic model is so popular. The model is designed to describe a probability, which is always some number between 0 and 1. In epidemiologic terms, such a probability gives the *risk* of an individual getting a disease.

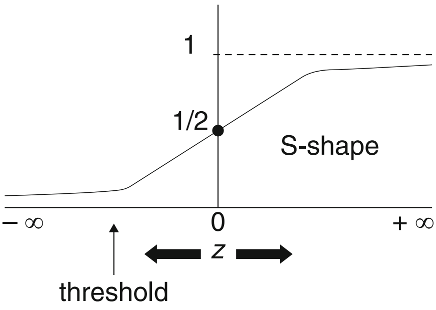
The *logistic model*, therefore, is set up to ensure that whatever estimate of risk we get, it will always be some number between 0 and 1. Thus, for the logistic model, we can never get a risk estimate either above 1 or below 0. This is not always true for other possible models, which is why the logistic model is often the first choice when a probability is to be estimated.

Shape:



Another reason why the logistic model is popular derives from the *shape* of the logistic function. As shown in the graph, it we start at $z = -\infty$ and move to the right, then as z increases, the value of $f(z)$ hovers close to zero for a while, then starts to increase dramatically toward 1, and finally levels off around 1 as z increases toward $+\infty$. The result is an elongated, S-shaped picture.

z = index of combined risk factors



The S-shape of the logistic function appeals to epidemiologists if the variable z is viewed as representing an index that combines contributions of several risk factors, and $f(z)$ represents the risk for a given value of z.

Then, the S-shape of $f(z)$ indicates that the effect of z on an individual's risk is minimal for low zs until some *threshold* is reached. The risk then rises rapidly over a certain range of intermediate z values and then remains extremely high around 1 once z gets large enough.

This *threshold* idea is thought by epidemiologists to apply to a variety of disease conditions. In other words, an S-shaped model is considered to be widely applicable for considering the multivariable nature of an epidemiologic research question.

SUMMARY

So, the logistic *model* is *popular* because the logistic *function*, on which the model is based, provides the following:

- Estimates that must lie in the range between zero and one
- An appealing S-shaped description of the combined effect of several risk factors on the risk for a disease.

III. The Logistic Model

Now, let us go from the logistic *function* to the *model*, which is our primary focus.

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

To obtain the logistic *model* from the logistic *function*, we write z as the linear sum α plus β_1 times X_1 plus β_2 times X_2, and so on to β_k times X_k, where the Xs are independent variables of interest and α and the β_i are constant terms representing unknown parameters.

$$z = \underbrace{\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}$$

$$f(z) = \frac{1}{1 + e^{-z}}$$

$$= \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}$$

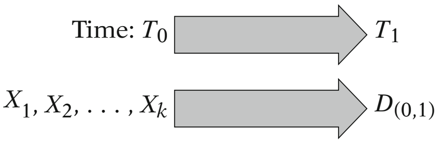
In essence, then, *z is an index that combines the Xs.*

We now substitute the linear sum expression for z in the right-hand side of the formula for $f(z)$ to get the expression $f(z)$ equals 1 over 1 plus e to minus the quantity α plus the sum of $\beta_i X_i$ for i ranging from 1 to k. Actually, to view this expression as a mathematical model, we must place it in an epidemiologic context.

Epidemiologic framework

$X_1, X_2, \ldots, X_k$ measured at T_0

The logistic model considers the following general *epidemiologic study framework*: We have observed independent variables X_1, X_2, and so on up to X_k on a group of subjects, for whom we have also determined disease status, as either 1 if “with disease” or 0 if “without disease”.



We wish to use this information to describe the probability that the disease will develop during a defined study period, say T_0 to T_1, in a disease-free individual with independent variable values X_1, X_2, up to X_k, which are measured at T_0.

$P(D = 1|X_1, X_2, \ldots, X_k)$

The probability being modeled can be denoted by the conditional probability statement $P(D{=}1 \mid X_1, X_2, \ldots, X_k)$.

DEFINITION

Logistic model:

$$P(D = 1|X_1, X_2, \ldots, X_k) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}$$

$$\uparrow \quad \uparrow$$

unknown parameters

The model is defined as *logistic* if the expression for the probability of developing the disease, given the Xs, is 1 over 1 plus e to minus the quantity α plus the sum from i equals 1 to k of β_i times X_i.

The terms α and β_i in this model represent *unknown parameters* that we need to estimate based on data obtained on the Xs and on D (disease outcome) for a group of subjects.

Thus, if we knew the parameters α and the β_i and we had determined the values of X_1 through X_k for a particular disease-free individual, we could use this formula to plug in these values and obtain the probability that this individual would develop the disease over some defined follow-up time interval.

NOTATION

$P(D = 1|X_1, X_2, \ldots, X_k)$

$= P(\mathbf{X})$

For notational convenience, we will denote the probability statement $P(D{=}1 |X_1, X_2, \ldots, X_k)$ as simply $P(\mathbf{X})$ where the *bold* $\mathbf{X}$ is a shortcut notation for the collection of variables X_1 through X_k.

Model formula:

$$\boxed{P(\mathbf{X}) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}}$$

Thus, the logistic model may be written as $P(\mathbf{X})$ equals 1 over 1 plus e to minus the quantity α plus the sum $\beta_i X_i$.

IV. Applying the Logistic Model Formula

EXAMPLE

$D = \text{CHD}_{(0,\ 1)}$

$X_1 = \text{CAT}_{(0,\ 1)}$

$X_2 = \text{AGE}_{\text{continuous}}$

$X_3 = \text{ECG}_{(0,\ 1)}$

$n = 609$ white males

9-year follow-up

$$\text{P}(\mathbf{X}) = \frac{1}{1 + e^{-(\alpha+\beta_1\text{CAT}+\beta_2\text{AGE}+\beta_3\text{ECG})}}$$

To illustrate the use of the logistic model, suppose the disease of interest is D equals CHD. Here CHD is coded 1 if a person has the disease and 0 if not.

We have three independent variables of interest: $X_1 = \text{CAT}$, $X_2 = \text{AGE}$, and $X_3 = \text{ECG}$. CAT stands for catecholamine level and is coded 1 if high and 0 if low, AGE is continuous, and ECG denotes electrocardiogram status and is coded 1 if abnormal and 0 if normal.

We have a data set of 609 white males on which we measured CAT, AGE, and ECG at the start of study. These people were then followed for 9 years to determine CHD status.

Suppose that in the analysis of this data set, we consider a logistic model given by the expression shown here.

DEFINITION
fit: use data to estimate

$\alpha, \beta_1, \beta_2, \beta_3$

We would like to *"fit"* this model; that is, we wish to use the data set to estimate the unknown parameters α, β_1, β_2, and β_3.

NOTATION
hat = ˆ

parameter $\Longleftrightarrow$ estimator

$\alpha\, \beta_1\, \beta_2 \qquad \hat{\alpha}\, \hat{\beta}_1\, \hat{\beta}_2$

Using common statistical notation, we distinguish the parameters from their estimators by putting a *hat* symbol on top of a parameter to denote its estimator. Thus, the estimators of interest here are α "hat," β_1 "hat," β_2 "hat," and β_3 "hat".

Method of estimation:
  maximum likelihood (ML) – see Chaps. 4 and 5

The method used to obtain these estimates is called *maximum likelihood* (ML). In two later chapters (Chaps. 4 and 5), we describe how the ML method works and how to test hypotheses and derive confidence intervals about model parameters.

EXAMPLE

$\hat{\alpha} = -3.911$

$\hat{\beta}_1 = 0.652$

$\hat{\beta}_2 = 0.029$

$\hat{\beta}_3 = 0.342$

Suppose the results of our model fitting yield the estimated parameters shown on the left.

EXAMPLE (continued)

$$\hat{P}(\mathbf{X}) = \frac{1}{1 + e^{-[-3.911+0.652(\text{CAT})+0.029(\text{AGE})+0.342(\text{ECG})]}}$$

Our fitted model thus becomes $\hat{P}(\mathbf{X})$ equals 1 over 1 plus e to minus the linear sum −3.911 plus 0.652 times CAT plus 0.029 times AGE plus 0.342 times ECG. We have replaced **P** by $\hat{P}(\mathbf{X})$ on the left-hand side of the formula because our estimated model will give us an estimated probability, not the exact probability.

☆ $\hat{P}(\mathbf{X}) = ?$

Suppose we want to use our fitted model, to obtain the predicted risk for a *certain individual*.

☆ CAT = ?
AGE = ?
ECG = ? → $\hat{P}(\mathbf{X})$ predicted risk

To do so, we would need to specify the values of the independent variables (CAT, AGE, ECG) for this individual and then plug these values into the formula for the fitted model to compute the estimated probability, $\hat{P}(\mathbf{X})$ for this individual. This estimate is often called a “predicted risk”, or simply “risk”.

☆ CAT = 1
AGE = 40
ECG = 0

To illustrate the calculation of a predicted risk, suppose we consider an individual with CAT = 1, AGE = 40, and ECG = 0.

$$\begin{aligned}\hat{P}(\mathbf{X}) &= \frac{1}{1 + e^{-\left[-3.911+0.652(1)+0.029(40)+0.342(0)\right]}}\\ &= \frac{1}{1 + e^{-(-2.101)}}\\ &= \frac{1}{1 + 8.173}\\ &= 0.1090, \text{ i.e., risk} \simeq 11\%\end{aligned}$$

Plugging these values into the fitted model gives us 1 over 1 plus e to minus the quantity −3.911 plus 0.652 times 1 plus 0.029 times 40 plus 0.342 times 0. This expression simplifies to 1 over 1 plus e to minus the quantity −2.101, which further reduces to 1 over 1 plus 8.173, which yields the value 0.1090.

Thus, for a person with CAT = 1, AGE = 40, and ECG = 0, the predicted risk obtained from the fitted model is 0.1090. That is, this person’s estimated risk is about 11%.

☆ CAT = 1
AGE = 40
ECG = 0 / ☆ CAT = 0
AGE = 40
ECG = 0

Here, for the same fitted model, we compare the predicted risk of a person with CAT = 1, AGE = 40, and ECG = 0 with that of a person with CAT = 0, AGE = 40, and ECG = 0.

$$\frac{\hat{P}_1(\mathbf{X})}{\hat{P}_0(\mathbf{X})} = \frac{0.1090}{0.0600}$$

11% risk / 6% risk

We previously computed the risk value of 0.1090 for the first person. The second probability is computed the same way, but this time we must replace CAT = 1 with CAT = 0. The predicted risk for this person turns out to be 0.0600. Thus, using the fitted model, the person with a high catecholamine level has an 11% *risk* for CHD, whereas the person with a low catecholamine level has a 6% *risk* for CHD over the period of follow-up of the study.

EXAMPLE

$$\frac{\hat{P}_1(\mathbf{X})}{\hat{P}_0(\mathbf{X})} = \frac{0.109}{0.060} = 1.82 \text{ risk ratio } (\widehat{\mathbf{RR}})$$

Note that, in this example, if we divide the predicted risk of the person with high catecholamine by that of the person with low catecholamine, we get a *risk ratio* estimate, denoted by $\widehat{\mathbf{RR}}$, of 1.82. Thus, using the fitted model, we find that the person with high CAT has almost twice the risk of the person with low CAT, assuming both persons are of AGE 40 and have no previous ECG abnormality.

- RR (direct method)

We have just seen that it is possible to use a logistic model to obtain a risk ratio estimate that compares two types of individuals. We will refer to the approach we have illustrated above as the *direct method* for estimating RR.

Conditions for RR (direct method):
✓ Follow-up study
✓ Specify all *X*s

Two conditions must be satisfied to estimate RR directly. First, we must have a *follow-up study* so that we can legitimately estimate individual risk. Second, for the two individuals being compared, we must *specify values for all the independent variables* in our fitted model to compute risk estimates for each individual.

- RR (indirect method):
  ✓ OR
  ✓ Assumptions

If either of the above conditions is not satisfied, then we cannot estimate RR directly. That is, if our study design is not a follow-up study *or* if some of the *X*s are not specified, we cannot estimate RR directly. Nevertheless, it may be possible to estimate RR *indirectly*. To do this, we must first compute an *odds ratio*, usually denoted as *OR*, and we must make some assumptions that we will describe shortly.

- OR: direct estimate from:
  ✓ Follow-up
  ✓ Case-control
  ✓ Cross-sectional

In fact, *the odds ratio* (OR), not the risk ratio (RR), is the only measure of association *directly estimated* from a logistic model (without requiring special assumptions), regardless of whether the study design is *follow-up, case- control*, or *cross-sectional*. To see how we can use the logistic model to get an odds ratio, we need to look more closely at some of the features of the model.

V. Study Design Issues

★ Follow-up study orientation

$$X_1, X_2, \ldots, X_k \Longrightarrow \boldsymbol{D}_{(0,1)}$$

An important feature of the logistic model is that it is defined with a *follow-up study orientation*. That is, as defined, this model describes the probability of developing a disease of interest expressed as a function of independent variables presumed to have been measured at the start of a fixed follow-up period. For this reason, it is natural to wonder whether the model can be applied to case-control or cross-sectional studies.

✓ *Case-control*
✓ *Cross-sectional*

The answer is *yes*: logistic regression can be applied to study designs other than follow-up.

Breslow and Day (1981)
Prentice and Pike (1979)

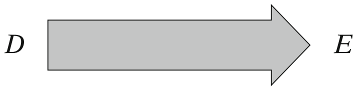
Robust conditions
Case-control studies

Robust conditions
Cross-sectional studies

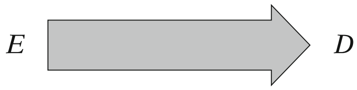
Two papers, one by *Breslow* and *Day* in 1981 and the other by *Prentice* and *Pike* in 1979 have identified certain *"robust" conditions* under which the logistic model can be used with case-control data. "Robust" means that the conditions required, which are quite complex mathematically and equally as complex to verify empirically, apply to a large number of data situations that actually occur.

The reasoning provided in these papers carries over to *cross-sectional studies* also, though this has not been explicitly demonstrated in the literature.

Case control:



Follow-up:



In terms of *case-control* studies, it has been shown that even though cases and controls are selected first, after which previous exposure status is determined, the analysis may proceed as if the selection process were the other way around, as in a follow-up study.

Treat case control like follow-up

In other words, even with a case-control design, one can pretend, when doing the analysis, that the dependent variable is disease outcome and the independent variables are exposure status plus any covariates of interest. When using a logistic model with a case-control design, you can treat the data as if it came from a follow-up study and still get a *valid* answer.

LIMITATION

case-control and
cross-sectional studies:

~~individual risk~~

✓ OR

Although logistic modeling is applicable to case-control and cross-sectional studies, there is one important *limitation* in the analysis of such studies. Whereas in follow-up studies, as we demonstrated earlier, a fitted logistic model can be used to predict the risk for an individual with specified independent variables, this model cannot be used to predict individual risk for case-control or cross-sectional studies. In fact, only estimates of *odds ratios* can be obtained for case-control and cross-sectional studies.

Simple Analysis

	$E = 1$	$E = 0$
$D = 1$	a	b
$D = 0$	c	d

The fact that only odds ratios, not individual risks, can be estimated from logistic modeling in case-control or cross-sectional studies is not surprising. This phenomenon is a carryover of a principle applied to simpler data analysis situations, in particular, to the simple analysis of a 2 × 2 table, as shown here.

Risk: only in follow-up
OR: case-control or cross-sectional

For a 2 × 2 table, *risk estimates* can be used *only* if the data derive from a follow-up study, whereas only *odds ratios* are appropriate if the data derive from a casecontrol or cross-sectional study.

$\widehat{\text{OR}} = ad/bc$

To explain this further, recall that for 2×2 tables, the odds ratio is calculated as $\widehat{\text{OR}}$ equals a times d over b times c, where $a, b, c,$ and d are the cell frequencies inside the table.

Case-control and cross-sectional studies:

$$= \frac{\hat{\text{P}}(E = 1 | D = 1) \big/ \hat{\text{P}}(E = 0 | D = 1)}{\hat{\text{P}}(E = 1 | D = 0) \big/ \hat{\text{P}}(E = 0 | D = 0)}$$

In case-control and cross-sectional studies, this OR formula can alternatively be written, as shown here, as a ratio involving probabilities for exposure status conditional on disease status.

$$\left.\begin{matrix}\hat{\text{P}}(E=1 \mid D=1) \\ \hat{\text{P}}(E=1 \mid D=0)\end{matrix}\right\} \text{P}(E|D) \text{ (general form)}$$

In this formula, for example, the term $\hat{\text{P}}(E = 1 | D = 1)$ is the estimated probability of being exposed, given that you are diseased. Similarly, the expression $\hat{\text{P}}(E = 1 | D = 0)$ is the estimated probability of being exposed given that you are not diseased. All the probabilities in this expression are of the general form $\text{P}(E \mid D)$.

Risk : $\text{P}(D \mid E)$

In contrast, in follow-up studies, formulae for risk estimates are of the form $\text{P}(D \mid E)$, in which the exposure and disease variables have been switched to the opposite side of the "given" sign.

$$\widehat{\text{RR}} = \frac{\hat{\text{P}}(D = 1 | E = 1)}{\hat{\text{P}}(D = 1 | E = 0)}$$

For example, the risk ratio formula for follow-up studies is shown here. Both the numerator and denominator in this expression are of the form $\text{P}(D \mid E)$.

Case-control or cross-sectional studies:

$\cancel{\mathrm{P}(D|E)}$

✓ $\mathrm{P}(E \mid D) \neq$ risk

Thus, in case-control or cross-sectional studies, risk estimates cannot be estimated because such estimates require conditional probabilities of the form $\mathrm{P}(D \mid E)$, whereas only estimates of the form $\mathrm{P}(E \mid D)$ are possible. This classic feature of a simple analysis also carries over to a logistic analysis.

$$\hat{\mathrm{P}}(\mathbf{X}) = \frac{1}{1+e^{-(\hat{\alpha}+\Sigma\hat{\beta}_i X_i)}}$$

estimates

There is a simple *mathematical explanation* for why predicted risks cannot be estimated using logistic regression for case-control studies. To see this, we consider the parameters α and the βs in the logistic model. To get a predicted risk $\hat{\mathrm{P}}(\mathbf{X})$ from fitting this model, we must obtain valid estimates of α and the βs, these estimates being denoted by "hats" over the parameters in the mathematical formula for the model.

Case control:

$\cancel{\hat{\alpha}} \Rightarrow \hat{\mathrm{P}}(\cancel{\mathbf{X}})$

When using logistic regression for case-control data, the parameter α cannot be validly estimated without knowing the sampling fraction of the population. Without having a "good" estimate of α, we cannot obtain a good estimate of the predicted risk $\hat{\mathrm{P}}(\mathbf{X})$ because $\hat{\alpha}$ is required for the computation.

Follow-up:

$\hat{\alpha} \Rightarrow \hat{\mathrm{P}}(\mathbf{X})$

In contrast, in follow-up studies, α can be estimated validly, and, thus, $\mathrm{P}(\mathbf{X})$ can also be estimated.

Case-control and cross-sectional:

✓$\hat{\beta}_i$, $\widehat{\mathrm{OR}}$

Now, although α cannot be estimated from a case-control or cross-sectional study, the βs can be estimated from such studies. As we shall see shortly, the βs provide information about odds ratios of interest. Thus, even though we cannot estimate α in such studies, and therefore cannot obtain predicted risks, we can, nevertheless, obtain estimated measures of association in terms of odds ratios.

EXAMPLE

Case-control Printout

Variable	Coefficient
Constant	$-4.50 = \hat{\alpha}$
X_1	$0.70 = \hat{\beta}_1$
X_2	$0.05 = \hat{\beta}_2$
X_3	$0.42 = \hat{\beta}_3$

Note that if a logistic model is fit to case-control data, most computer packages carrying out this task will provide numbers corresponding to all parameters involved in the model, including α. This is illustrated here with some fictitious numbers involving three variables, X_1, X_2, and X_3. These numbers include a value corresponding to α, namely, -4.5, which corresponds to the constant on the list.

EXAMPLE (repeated)

Case-control Printout

Variable	Coefficient
Constant	$-4.50 = \hat{\alpha}$
X_1	$0.70 = \hat{\beta}_1$
X_2	$0.05 = \hat{\beta}_2$
X_3	$0.42 = \hat{\beta}_3$

$\hat{\alpha}$ not a valid estimate of α

However, according to mathematical theory, the value provided for the constant does not really estimate α. In fact, this value estimates some other parameter of no real interest. Therefore, an investigator should be forewarned that, even though the computer will print out a number corresponding to the constant α, the number will not be an appropriate estimate of α in case-control or cross-sectional studies.

SUMMARY

	Logistic Model	$\hat{P}(\mathbf{X})$	OR
Follow-up	✓	✓	✓
Case-control	✓	X	✓
Cross-sectional	✓	X	✓

We have described that the logistic model can be applied to case-control and cross-sectional data, even though it is intended for a follow-up design. When using case-control or cross-sectional data, however, a key limitation is that you cannot estimate risks like $\hat{P}(\mathbf{X})$, even though you can still obtain odds ratios. This limitation is not extremely severe if the goal of the study is to obtain a valid estimate of an exposure–disease association in terms of an odds ratio.

VI. Risk Ratios vs. Odds Ratios

OR
vs. ? follow-up study
RR

The use of an odds ratio estimate may still be of some concern, particularly when the study is a follow-up study. In follow-up studies, it is commonly preferred to estimate a risk ratio rather than an odds ratio.

EXAMPLE

$$\widehat{\mathrm{RR}} = \frac{\hat{P}(\mathrm{CHD}=1\,|\,\mathrm{CAT}=1,\ \mathrm{AGE}=40,\ \mathrm{ECG}=0)}{\hat{P}(\mathrm{CHD}=1\,|\,\mathrm{CAT}=0,\ \mathrm{AGE}=40,\ \mathrm{ECG}=0)}$$

Model:

$$P(\mathbf{X}) = \frac{1}{1+e^{-(\alpha+\beta_1\,\mathrm{CAT}+\beta_2\,\mathrm{AGE}+\beta_3\mathrm{ECG})}}$$

We previously illustrated that a risk ratio can be estimated for follow-up data provided all the independent variables in the fitted model are specified. In the example, we showed that we could estimate the risk ratio for CHD by comparing high catecholamine persons (that is, those with CAT = 1) to low catecholamine persons (those with CAT = 0), given that both persons were 40 years old and had no previous ECG abnormality. Here, we have specified values for all the independent variables in our model, namely, CAT, AGE, and ECG, for the two types of persons we are comparing.

EXAMPLE (continued)

$$\widehat{\text{RR}} = \frac{\hat{\text{P}}(\text{CHD} = 1 \mid \text{CAT} = 1, \text{AGE} = 40, \text{ECG} = 0)}{\hat{\text{P}}(\text{CHD} = 1 \mid \text{CAT} = 0, \text{AGE} = 40, \text{ECG} = 0)}$$

AGE uspecified but fixed

ECG unspecified but fixed

Nevertheless, it is more common to obtain an estimate of RR or OR without explicitly specifying the control variables. In our example, we want to compare high CAT with low CAT persons keeping the control variables like AGE and ECG fixed but unspecified. In other words, the question is typically asked: What is the effect of the CAT variable controlling for AGE and ECG, considering persons who have the same AGE and ECG, regardless of the values of these two variables?

Control variables unspecified:

$\widehat{\text{OR}}$ directly

$\widehat{\text{RR}}$ indirectly
provided $\widehat{\text{OR}} \approx \widehat{\text{RR}}$

When the control variables are generally considered to be fixed, but *unspecified*, as in the last example, we can use logistic regression to obtain an estimate of the OR *directly*, but we cannot estimate the RR. We can, however, obtain a RR *indirectly* if we can justify using the *rare disease assumption*, which assumes that the disease is sufficiently "rare" to allow the OR to provide a close approximation to the RR.

Rare disease	OR	RR (or PR)
Yes	√	√
No	√	Other

Other √ Log-binomial model
Poisson model
COPY method

If we cannot invoke the rare disease assumption, several alternative methods for estimating an adjusted RR (or prevalence ratio, PR) from logistic modeling have been proposed in the recent literature. These include "standardization" (Wilcosky and Chambless, 1985 and Flanders and Rhodes, 1987); a "case-cohort model" (Schouten et al., 1993); a "log-binomial model (Wacholder, 1986 and Skov et al., 1998); a "Poisson regression model" (McNutt et al., 2003 and Barros and Hirakata, 2003); and a "COPY method" (Deddens and Petersen, 2008).

The latter paper reviews all previous approaches. They conclude that a log-binomial model should be preferred when estimating RR or PR in a study with a common outcome. However, if the log-binomial model does not converge, they recommend using either the COPY method or the robust Poisson method. For further details, see the above references.

VII. Logit Transformation

OR: Derive and Compute

Having described why the odds ratio is the primary parameter estimated when fitting a logistic regression model, we now explain how an odds ratio is derived and computed from the logistic model.

Logit

To begin the description of the odds ratio in logistic regression, we present an alternative way to write the logistic model, called the *logit form* of the model. To get the *logit* from the logistic model, we make a transformation of the model.

$$\text{logit P}(\mathbf{X}) = \ln_e\left[\frac{\text{P}(\mathbf{X})}{1-\text{P}(\mathbf{X})}\right],$$

where

$$\text{P}(\mathbf{X}) = \frac{1}{1+e^{-(\alpha+\sum\beta_i X_i)}}$$

The *logit transformation*, denoted as *logit* P(**X**), is given by the natural log (i.e., to the base e) of the quantity P(**X**) divided by one minus P(**X**), where P(**X**) denotes the logistic model as previously defined.

This transformation allows us to compute a number, called **logit** P(**X**), for an individual with independent variables given by **X**. We do so by:

(1) $\text{P}(\mathbf{X})$

(2) $1-\text{P}(\mathbf{X})$

(3) $\dfrac{\text{P}(\mathbf{X})}{1-\text{P}(\mathbf{X})}$

(4) $\ln_e\left[\dfrac{\text{P}(\mathbf{X})}{1-\text{P}(\mathbf{X})}\right]$

(1) computing P(**X**) and
(2) 1 minus P(**X**) separately, then
(3) dividing one by the other, and finally
(4) taking the natural log of the ratio.

EXAMPLE

(1) $\text{P}(\mathbf{X}) = 0.110$

(2) $1-\text{P}(\mathbf{X}) = 0.890$

(3) $\dfrac{\text{P}(\mathbf{X})}{1-\text{P}(\mathbf{X})} = \dfrac{0.110}{0.890} = 0.123$

(4) $\ln_e\left[\frac{\text{P}(\mathbf{X})}{1-\text{P}(\mathbf{X})}\right] = \ln(0.123) = -2.096$

i.e., logit (0.110) = −2.096

For example, if P(**X**) is 0.110, then

1 minus P(**X**) is 0.890,

the ratio of the two quantities is 0.123,

and the log of the ratio is −2.096.

That is, the *logit* of 0.110 is −2.096.

$$\text{logit P}(\mathbf{X}) = \ln_e\left[\frac{\text{P}(\mathbf{X})}{1-\text{P}(\mathbf{X})}\right] = ?$$

$$\text{P}(\mathbf{X}) = \frac{1}{1+e^{-(\alpha+\sum\beta_i X_i)}}$$

Now we might ask, *what general formula do we get when we plug the logistic model form into the logit function? What kind of interpretation can we give to this formula? How does this relate to an odds ratio?*

Let us consider the formula for the logit function. We start with P(**X**), which is 1 over 1 plus e to minus the quantity α plus the sum of the $\beta_i X_i$.

$$1 - \mathrm{P}(\mathbf{X}) = 1 - \frac{1}{1 + e^{-\left(\alpha + \sum \beta_i X_i\right)}} = \frac{e^{-\left(\alpha + \sum \beta_i X_i\right)}}{1 + e^{-\left(\alpha + \sum \beta_i X_i\right)}}$$

Also, using some algebra, we can write 1 − P(**X**) as:

e to minus the quantity α plus the sum of β_iX$_i$ divided by one over 1 plus e to minus α plus the sum of the β_iX$_i$.

$$\frac{\mathrm{P}(\mathbf{X})}{1 - \mathrm{P}(\mathbf{X})} = \frac{\dfrac{1}{\cancel{1 + e^{-(\alpha + \sum \beta_i X_i)}}}}{\dfrac{e^{-(\alpha + \sum \beta_i X_i)}}{\cancel{1 + e^{-(\alpha + \sum \beta_i X_i)}}}}$$

If we divide P(**X**) by 1 − P(**X**), then the denominators cancel out,

$$= e^{\left(\alpha + \sum \beta_i X_i\right)}$$

and we obtain e to the quantity α plus the sum of the β_iX$_i$.

$$\ln_e\left[\frac{\mathrm{P}(\mathbf{X})}{1 - \mathrm{P}(\mathbf{X})}\right] = \ln_e\left[e^{\left(\alpha + \sum \beta_i X_i\right)}\right] = \underbrace{\left(\alpha + \sum \beta_i X_i\right)}_{\text{linear sum}}$$

We then compute the natural log of the formula just derived to obtain:

the linear sum α plus the sum of β_iX$_i$.

Thus, the *logit* of P(**X**) simplifies to the *linear sum* found in the denominator of the formula for P(**X**).

Logit form: $$\boxed{\begin{array}{l}\text{logit } \mathrm{P}(\mathrm{X}) = \alpha + \sum \beta_i X_i, \\ \quad \text{where} \\ \mathrm{P}(\mathbf{X}) = \dfrac{1}{1 + e^{-\left(\alpha + \sum \beta_i X_i\right)}}\end{array}}$$

For the sake of convenience, many authors describe the logistic model in its logit form rather than in its original form as P(**X**). Thus, when someone describes a model as *logit* P(**X**) equal to a linear sum, we should recognize that a logistic model is being used.

$$\text{logit P(X)} \xrightarrow{\quad ? \quad} \text{OR}$$

$$\frac{\mathrm{P}(\mathbf{X})}{1 - \mathrm{P}(\mathbf{X})} = \text{odds for individual } X$$

Now, having defined and expressed the formula for the logit form of the logistic model, we ask, *where does the odds ratio come in*? As a preliminary step to answering this question, we first look more closely at the definition of the logit function. In particular, the quantity P(**X**) divided by 1 − P(**X**), whose log value gives the *logit*, describes the *odds* for developing the disease for a person with independent variables specified by **X**.

$$\text{odds} = \frac{P}{1 - P}$$

In its simplest form, an *odds* is the ratio of the probability that some event will occur over the probability that the same event will not occur. The formula for an odds is, therefore, of the form P divided by 1−P, where P denotes the probability of the event of interest.

EXAMPLE

$P = 0.25$

$$\text{odds} = \frac{P}{1-P} = \frac{0.25}{0.75} = \frac{1}{3}$$

$1 \leftarrow$ event occurs
$3 \leftarrow$ event does not occur

3 to 1 event will not happen

For example, if P equals 0.25, then $1 - P$, the probability of the opposite event, is 0.75 and the *odds* is 0.25 over 0.75, or one-third.

An *odds* of one-third can be interpreted to mean that the probability of the event occurring is one-third the probability of the event not occurring. Alternatively, we can state that the *odds* are 3 *to* 1 that the event will not happen.

The expression P(**X**) divided by 1 − P(**X**) has essentially the same interpretation as P over $1 - P$, which ignores **X**.

$$\text{odds}: \left[\underbrace{\frac{\mathrm{P}(\mathbf{X})}{1-\mathrm{P}(\mathbf{X})}}_{\text{describes risk in logistic model for individual } \mathbf{X}}\right] \text{vs.} \frac{P}{1-P}$$

The main difference between the two formulae is that the expression with the **X** is more specific. That is, the formula with **X** assumes that the probabilities describe the risk for developing a disease, that this risk is determined by a logistic model involving independent variables summarized by **X**, and that we are interested in the odds associated with a particular specification of **X**.

$$\begin{aligned}\text{logit}\,\mathrm{P}(\mathbf{X}) &= \ln_e\left[\frac{\mathrm{P}(\mathbf{X})}{1-\mathrm{P}(\mathbf{X})}\right]\\ &= \text{log odds for individual } \mathbf{X}\\ &= \alpha + \sum \beta_i X_i\end{aligned}$$

Thus, the logit form of the logistic model, shown again here, gives an expression for the *log odds* of developing the disease for an individual with a specific set of Xs.

And, mathematically, this expression equals α plus the sum of the $\beta_i X_i$.

EXAMPLE

all $X_i = 0$: logit P(**X**) = ?

$$\text{logit P}(\mathbf{X}) = \alpha + \overbrace{\Sigma\beta_i X_i}^{0}$$

$$\text{logit P}(\mathbf{X}) \Rightarrow \alpha$$

As a simple example, consider what the *logit* becomes when all the Xs are 0. To compute this, we need to work with the mathematical formula, which involves the unknown parameters and the Xs.

If we plug in 0 for all the Xs in the formula, we find that the logit of P(**X**) reduces simply to α.

Because we have already seen that any logit can be described in terms of an *odds*, we can interpret this result to give some meaning to the parameter α.

INTERPRETATION

(1) α = log odds for individual with all $X_i = 0$

One interpretation is that α gives the *log odds* for a person with zero values for all Xs.

EXAMPLE (continued)

(2) α = log of background odds

A second interpretation is that α gives the *log* of the *background*, or *baseline, odds*.

LIMITATION OF (1)

All $X_\mathrm{i} = 0$ for any individual?

The first interpretation for α, which considers it as the *log odds* for a person with 0 values for all *X*s, has a serious limitation: There may not be any person in the population of interest with zero values on all the *X*s.

AGE $\neq 0$
WEIGHT $\neq 0$

For example, no subject could have zero values for naturally occurring variables, like age or weight. Thus, it would not make sense to talk of a person with zero values for all *X*s.

DEFINITION OF (2)

background odds: ignores all *X*s

The second interpretation for α is more appealing: to describe it as the *log* of the *background*, or *baseline, odds*.

By background odds, we mean the odds that would result for a logistic model without any *X*s at all.

$$\text{model}: \mathrm{P}(\mathbf{X}) = \frac{1}{1 + e^{-\alpha}}$$

The form of such a model is 1 over 1 plus e to minus α. We might be interested in this model to obtain a baseline risk or odds estimate that ignores all possible predictor variables. Such an estimate can serve as a starting point for comparing other estimates of risk or odds when one or more *X*s are considered.

α ✓
β_i?

Because we have given an interpretation to α, can we also give an interpretation to β_i? Yes, we can, in terms of either *odds* or *odds ratios*. We will turn to odds ratios shortly.

$$\underset{\text{fixed}}{X_1, X_2}, \ldots, \underset{\text{varies}}{X_i}, \ldots, \underset{\text{fixed}}{X_k}$$

With regard to the odds, we need to consider what happens to the logit when only one of the *X*s varies while keeping the others fixed.

EXAMPLE

CAT changes from 0 to 1;

$$\underbrace{\text{AGE} = 40, \text{ECG} = 0}_{\text{fixed}}$$

For example, if our *X*s are CAT, AGE, and ECG, we might ask what happens to the logit when CAT changes from 0 to 1, given an AGE of 40 and an ECG of 0.

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1\text{CAT} + \beta_2\text{AGE} + \beta_3\text{ECG}$$

To answer this question, we write the model in *logit form* as $\alpha + \beta_1\text{CAT} + \beta_2\text{AGE} + \beta_3\text{ECG}$.

EXAMPLE (continued)

(1) CAT = 1, AGE = 40, ECG = 0

$$\text{logit } P(\mathbf{X}) = \alpha + \beta_1 1 + \beta_2 40 + \beta_3 0 = \boxed{\alpha + \beta_1 + 40\beta_2}$$

The first expression below this model shows that when CAT = 1, AGE = 40, and ECG = 0, this logit reduces to $\alpha + \beta_1 + 40\beta_2$.

(2) CAT = 0, AGE = 40, ECG = 0

$$\text{logit } P(\mathbf{X}) = \alpha + \beta_1 0 + \beta_2 40 + \beta_3 0 = \boxed{\alpha + 40\beta_2}$$

The second expression shows that when CAT = 0, but AGE and ECG remain fixed at 40 and 0, respectively, the logit reduces to $\alpha + 40\,\beta_2$.

$$\text{logit } P_1(\mathbf{X}) - \text{logit } P_0(\mathbf{X}) = (\alpha + \beta_1 + 40\beta_2) - (\alpha + 40\beta_2) = \boxed{\beta_1}$$

If we subtract the *logit for CAT* = 0 from the *logit for CAT* = 1, after a little arithmetic, we find that the difference is β_1, the coefficient of the variable CAT.

NOTATION

$\triangle$ = change

$\beta_1 = \triangle$ logit

$= \triangle$ log odds

when $\triangle$ CAT = 1
AGE and ECG fixed

Thus, letting the symbol $\triangle$ denote change, we see that β_1 represents the change in the logit that would result from a unit change in CAT, when the other variables are fixed.

An equivalent explanation is that β_1 represents the *change in the log odds that would result from a one unit change* in the variable CAT when the other variables are fixed. These two statements are equivalent because, by definition, a *logit* is a *log odds*, so that the difference between two logits is the same as the difference between two log odds.

$$\text{logit } P(\mathbf{X}) = \alpha + \textstyle\sum \beta_i X_i$$

More generally, using the logit expression, if we focus on any coefficient, say β_L, for $i = L$, we can provide the following interpretation:

$i = L$:

$$\boxed{\beta_L = \triangle \ln(\text{odds})}$$

when $\triangle X_L = 1$, other *X*s fixed

β_L represents the change in the log odds that would result from a one unit change in the variable X_L, when all other *X*s are fixed.

SUMMARY

logit P(**X**)

α = background log odds

β_i = change in log odds

In summary, by looking closely at the expression for the logit function, we provide some interpretation for the parameters α and β_i in terms of odds, actually *log odds*.

Now, how can we use this information about logits to obtain an *odds ratio*, rather than an odds? After all, we are typically interested in measures of association, like odds ratios, when we carry out epidemiologic research.

VIII. Derivation of OR Formula

$$\text{OR} = \frac{\text{odds}_1}{\text{odds}_0}$$

Any *odds ratio*, by definition, is a ratio of two odds, written here as $odds_1$ divided by $odds_0$, in which the subscripts indicate two individuals or two groups of individuals being compared.

EXAMPLE

(1) CAT = 1, AGE = 40, ECG = 0

(0) CAT = 0, AGE = 40, ECG = 0

Now we give an example of an odds ratio in which we compare two groups, called group 1 and group 0. Using our CHD example involving independent variables CAT, AGE, and ECG, group 1 might denote persons with CAT = 1, AGE = 40, and ECG = 0, whereas group 0 might denote persons with CAT = 0, AGE = 40, and ECG = 0.

$\mathbf{X} = (X_1, X_2, \ldots, X_\mathrm{k})$

More generally, when we describe an odds ratio, the two groups being compared can be defined in terms of the bold **X** symbol, which denotes a general collection of X variables, from 1 to k.

(1) $\mathbf{X}_1 = (X_{11}, X_{12}, \ldots, X_{1k})$

(0) $\mathbf{X}_0 = (X_{01}, X_{02}, \ldots, X_{0k})$

Let $\mathbf{X}_1$ denote the collection of Xs that specify group 1 and let $\mathbf{X}_0$ denote the collection of Xs that specify group 0.

EXAMPLE

$\mathbf{X} = (\text{CAT, AGE, ECG})$

(1) $\mathbf{X}_1 = (\text{CAT} = 1, \text{AGE} = 40, \text{ECG} = 0)$

(0) $\mathbf{X}_0 = (\text{CAT} = 0, \text{AGE} = 40, \text{ECG} = 0)$

In our example, then, k, the number of variables, equals 3, and

X is the collection of variables CAT, AGE, and ECG,

$\mathbf{X}_1$ corresponds to CAT = 1, AGE = 40, and ECG = 0, whereas

$\mathbf{X}_0$ corresponds to CAT = 0, AGE = 40, and ECG = 0.

NOTATION

$$\text{OR}_{\mathbf{X}_1, \mathbf{X}_0} = \frac{\text{odds for } \mathbf{X}_1}{\text{odds for } \mathbf{X}_0}$$

Notationally, to distinguish the two groups $\mathbf{X}_1$ and $\mathbf{X}_0$ in an *odds ratio*, we can write $\text{OR}_{\mathbf{X}_1, \mathbf{X}_0}$ equals the *odds* for $\mathbf{X}_1$ *divided by* the *odds* for $\mathbf{X}_0$.

We will now apply the logistic model to this expression to obtain a general odds ratio formula involving the logistic model parameters.

$$P(\mathbf{X}) = \frac{1}{1 + e^{-\left(\alpha + \sum \beta_i X_i\right)}}$$

Given a logistic model of the general form P(**X**),

$$(1)\ \text{odds}: \frac{P(\mathbf{X}_1)}{1 - P(\mathbf{X}_1)}$$

we can write the *odds* for *group* 1 as P($\mathbf{X}_1$) divided by 1 − P($\mathbf{X}_1$)

$$(0)\ \text{odds}: \frac{P(\mathbf{X}_0)}{1 - P(\mathbf{X}_0)}$$

and the *odds* for *group* 0 as P($\mathbf{X}_0$) divided by 1 − P($\mathbf{X}_0$).

$$\frac{\text{odds for } \mathbf{X}_1}{\text{odds for } \mathbf{X}_0} = \frac{\dfrac{P(\mathbf{X}_1)}{1-P(\mathbf{X}_1)}}{\dfrac{P(\mathbf{X}_0)}{1-P(\mathbf{X}_0)}} = \text{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0}$$

To get an odds ratio, we then divide the first odds by the second odds. The result is an expression for the odds ratio written in terms of the two risks P($\mathbf{X}_1$) and P($\mathbf{X}_0$), that is, P($\mathbf{X}_1$) over 1 − P($\mathbf{X}_1$) divided by P($\mathbf{X}_0$) over 1 − P($\mathbf{X}_0$).

We denote this ratio as *ROR*, for *risk odds ratio*, as the probabilities in the odds ratio are all defined as risks. However, we still do not have a convenient formula.

$$\text{ROR} = \frac{\dfrac{P(\mathbf{X}_1)}{1 - P(\mathbf{X}_1)}}{\dfrac{P(\mathbf{X}_0)}{1 - P(\mathbf{X}_0)}} \qquad P(\mathbf{X}) = \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}}$$

$$(1)\ \frac{P(\mathbf{X}_1)}{1 - P(\mathbf{X}_1)} = e^{(\alpha + \sum \beta_i X_{1i})}$$

$$(0)\ \frac{P(\mathbf{X}_0)}{1 - P(\mathbf{X}_0)} = e^{(\alpha + \sum \beta_i X_{0i})}$$

Now, to obtain a convenient computational formula, we can substitute the mathematical expression 1 over 1 plus e to minus the quantity ($\alpha + \sum \beta_i X_i$) for P(**X**) into the *risk odds ratio* formula above.

For group 1, the *odds* P($\mathbf{X}_1$) over 1 − P($\mathbf{X}_1$) reduces algebraically to e to the linear sum α plus the sum of β_i times X_{1i}, where X_{1i} denotes the value of the variable X_i for group 1.

Similarly, the odds for group 0 reduces to e to the linear sum α plus the sum of β_i times X_{0i}, where X_{0i} denotes the value of variable X_i for group 0.

$$\text{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} = \frac{\text{odds for } \mathbf{X}_1}{\text{odds for } \mathbf{X}_0} = \frac{e^{\left(\alpha + \sum \beta_i X_{1i}\right)}}{e^{\left(\alpha + \sum \beta_i X_{0i}\right)}}$$

To obtain the *ROR*, we now substitute in the numerator and denominator the exponential quantities just derived to obtain e to the group 1 linear sum divided by e to the group 0 linear sum.

$$\text{Algebraic theory}: \frac{e^a}{e^b} = e^{a-b}$$

$$a = \alpha + \Sigma \beta_1 X_{1i}, \quad b = \alpha + \Sigma \beta_i X_{0i}$$

The above expression is of the form e to the *a* divided by e to the *b*, where *a* and *b* are linear sums for groups 1 and 0, respectively. From algebraic theory, it then follows that this ratio of two exponentials is equivalent to e to the difference in exponents, or e to the *a* minus *b*.

$$\text{ROR} = e^{(\alpha+\sum\beta_i X_{1i})-(\alpha+\sum\beta_i X_{0i})}$$

We then find that the *ROR* equals e to the difference between the two linear sums.

$$= e^{\left[\alpha-\alpha+\sum\beta_i(X_{1i}-X_{0i})\right]}$$

In computing this difference, the αs cancel out and the β_is can be factored for the *i*th variable.

$$= e^{\sum\beta_i(X_{1i}-X_{0i})}$$

Thus, the expression for *ROR* simplifies to the quantity e to the sum β_i times the difference between X_{1i} and X_{0i}.

- $$\boxed{\text{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} = e^{\sum\limits_{i=1}^{k}\beta_i(X_{1i}-X_{0i})}}$$

We thus have a general exponential formula for the risk odds ratio from a logistic model comparing any two groups of individuals, as specified in terms of $\boldsymbol{X}_1$ and $\boldsymbol{X}_0$. Note that the formula involves the β_is but not α.

$$\boxed{e^{a+b} = e^a \times e^b}$$

We can give an equivalent alternative to our ROR formula by using the algebraic rule that says that the exponential of a sum is the same as the product of the exponentials of each term in the sum. That is, e to the *a* plus *b* equals e to the *a* times e to the *b*.

$$e^{\sum\limits_{i=1}^{k} z_i} = e^{z_1} \times e^{z_2} \times \cdots e^{z_k}$$

More generally, e to the sum of z_i equals the product of e to the z_i over all *i*, where the z_i's denote any set of values.

NOTATION

$$= \prod_{i=1}^{k} e^{z_i}$$

$$z_i = \beta_i(X_{1i}-X_{0i})$$

We can alternatively write this expression using the product symbol Π, where Π is a mathematical notation which denotes the product of a collection of terms.

Thus, using algebraic theory and letting z_i correspond to the term β_i times $(X_{1i} - X_{0i})$,

- $$\boxed{\text{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} = \prod_{i=1}^{k} e^{\beta_i(X_{1i}-X_{0i})}}$$

we obtain the *alternative formula* for *ROR* as the product from $i = 1$ to *k* of e to the β_i times the difference $(X_{1i} - X_{0i})$.

$$\prod_{i=1}^{k} e^{\beta_i(X_{1i}-X_{0i})}$$

That is, Π of e to the β_i times $(X_{1i} - X_{0i})$ equals e to the β_1 times $(X_{11} - X_{01})$ multiplied by e to the β_2 times $(X_{12} - X_{02})$ multiplied by additional terms, the final term

$$= e^{\beta_1(X_{11}-X_{01})} e^{\beta_2(X_{12}-X_{02})} \ldots e^{\beta_k(X_{1k}-X_{0k})}$$

being e to the β_{k} times $(X_{1k} - X_{0k})$.

$$\mathrm{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} = \prod_{i=1}^{k} e^{\beta_i(X_{1i}-X_{0i})}$$

- Multiplicative

The *product formula* for the *ROR*, shown again here, gives us an interpretation about how each variable in a logistic model contributes to the odds ratio.

In particular, we can see that each of the variables X_i contributes jointly to the odds ratio in a *multiplicative* way.

For example, if

EXAMPLE

$e^{\beta_2(X_{12}-X_{02})} = 3$

$e^{\beta_5(X_{15}-X_{05})} = 4$

$3 \times 4 = 12$

e to the β_i times $(X_{1i} - X_{0i})$ is

3 for variable 2 and

4 for variable 5,

then the joint contribution of these two variables to the odds ratio is **3** × **4**, or **12**.

Logistic model ⇒ multiplicative OR formula

Thus, the product or Π formula for *ROR* tells us that, when the logistic model is used, the contribution of the variables to the odds ratio is *multiplicative*.

Other models ⇒ other OR formulae

A model different from the logistic model, depending on its form, might imply a different (e.g., an additive) contribution of variables to the odds ratio. An investigator not willing to allow a multiplicative relationship may, therefore, wish to consider other models or other OR formulae. Other such choices are beyond the scope of this presentation.

IX. Example of OR Computation

$$\mathrm{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} = e^{\sum_{i=1}^{k} \beta_i(X_{1i}-X_{0i})}$$

Given the choice of a logistic model, the version of the formula for the *ROR*, shown here as the exponential of a sum, is the most useful for computational purposes.

EXAMPLE

$\mathbf{X}$ = (CAT, AGE, ECG)
(1) CAT = 1, AGE = 40, ECG = 0
(0) CAT = 0, AGE = 40, ECG = 0

For example, suppose the **X**s are CAT, AGE, and ECG, as in our earlier examples.

Also suppose, as before, that we wish to obtain an expression for the odds ratio that compares the following two groups: *group 1* with CAT = 1, AGE = 40, and ECG = 0, and *group 0* with CAT = 0, AGE = 40, and ECG = 0.

$\mathbf{X}_1$ = (CAT = 1, AGE = 40, ECG = 0)

For this situation, we let $\mathbf{X}_1$ be specified by CAT = 1, AGE = 40, and ECG = 0,

EXAMPLE (continued)

$\mathbf{X}_0 = (\text{CAT} = 0, \text{AGE} = 40, \text{ECG} = 0)$

and let $\mathbf{X}_0$ be specified by CAT $= 0$, AGE $= 40$, and ECG $= 0$.

$\text{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} = e^{\sum\limits_{i=1}^{k} \beta_i (X_{1i} - X_{0i})}$

Starting with the general formula for the *ROR*, we then substitute the values for the $\mathbf{X}_1$ and $\mathbf{X}_0$ variables in the formula.

$= e^{\beta_1(1-0)+\beta_2(40-40)+\beta_3(0-0)}$

We then obtain *ROR* equals e to the β_1 times $(1 - 0)$ plus β_2 times $(40 - 40)$ plus β_3 times $(0 - 0)$.

$= e^{\beta_1+0+0}$

The last two terms reduce to 0,

$= e^{\beta_1} \longleftarrow$ coefficient of CAT in
logit $\text{P}(\mathbf{X}) = \alpha + \beta_1\text{CAT} + \beta_2\text{AGE} + \beta_3\text{ECG}$

so that our final expression for the *odds ratio* is e to the β_1, where β_1 is the coefficient of the variable CAT.

$\text{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} = e^{\beta_1}$

(1) CAT = 1, AGE = 40, ECG = 0
(0) CAT = 0, AGE = 40, ECG = 0

Thus, for our example, even though the model involves the three variables CAT, ECG, and AGE, the odds ratio expression comparing the two groups involves only the parameter involving the variable CAT. Notice that of the three variables in the model, the variable CAT is the only variable whose value is different in groups 1 and 0. In both groups, the value for AGE is 40 and the value for ECG is 0.

$\text{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} = e^{\beta_1}$
$\quad = \text{an ``adjusted'' OR}$

AGE and ECG:

- Fixed
- Same
- Control variables

The formula e to the β_1 may be interpreted, in the context of this example, as an *adjusted odds ratio*. This is because we have derived this expression from a logistic model containing two other variables, namely, AGE and ECG, in addition to the variable CAT. Furthermore, we have fixed the values of these other two variables to be the same for each group. Thus, e to β_1 gives an odds ratio for the effect of the CAT variable *adjusted* for AGE and ECG, where the latter two variables are being treated as *control variables*.

e^{β_1}: population ROR

The expression e to the β_1 denotes a population odds ratio parameter because the term β_1 is itself an unknown population parameter.

$e^{\hat{\beta}_1}$: estimated ROR

An estimate of this population odds ratio would be denoted by e to the $\hat{\beta}_1$. This term, $\hat{\beta}_1$, denotes an *estimate* of β_1 obtained by using some computer package to fit the logistic model to a set of data.

X. Special Case for (0, 1) Variables

Adjusted OR = e^{β}
where β = coefficient of (0, 1) variable

Our example illustrates an important special case of the general odds ratio formula for logistic regression that applies to (0, 1) variables. That is, an *adjusted odds ratio* can be obtained by exponentiating the coefficient of a (0, 1) variable in the model.

EXAMPLE

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1 \boxed{\text{CAT}} + \beta_2\text{AGE} + \beta_3\text{ECG}$$

adjusted

In our example, that variable is CAT, and the other two variables, AGE and ECG, are the ones for which we adjusted.

$X_i(0, 1)$: adj. ROR = e^{β_i}

$\boxed{\text{controlling}}$ for other Xs

More generally, if the variable of interest is X_i, a (0, 1) variable, then e to the β_i, where β_i is the coefficient of X_i, gives an adjusted odds ratio involving the effect of X_i adjusted or controlling for the remaining X variables in the model.

EXAMPLE

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1\text{CAT} + \beta_2\text{AGE} + \beta_3 \boxed{\text{ECG}}$$

adjusted

ECG (0, 1): adj. ROR = e^{β_3}

controlling for CAT and AGE

Suppose, for example, our focus had been on *ECG*, also a (0, 1) variable, instead of on CAT in a logistic model involving the same variables CAT, AGE, and ECG.

Then e to the β_3, where β_3 is the coefficient of ECG, would give the adjusted odds ratio for the effect of ECG, controlling for CAT and AGE.

SUMMARY

X_i is $(0, 1)$: ROR = e^{β_i}

General OR formula:

$$\text{ROR} = e^{\sum\limits_{i=1}^{k} \beta_i (X_{1i} - X_{0i})}$$

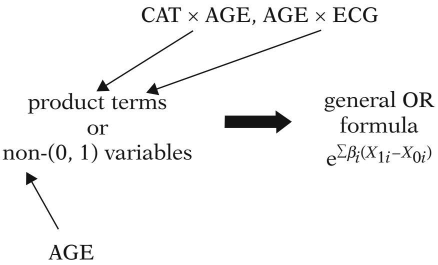
Thus, we can obtain an adjusted odds ratio for each (0, 1) variable in the logistic model by exponentiating the coefficient corresponding to that variable. This formula is much simpler than the general formula for ROR described earlier.

EXAMPLE

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1\text{CAT} + \beta_2\text{AGE} + \beta_3\text{ECG}$$

main effect variables

Note, however, that the example we have considered involves only *main effect variables*, like CAT, AGE and ECG, and that the model does not contain product terms like CAT × AGE or AGE × ECG.



CAT × AGE, AGE × ECG

product terms or non-(0, 1) variables

AGE

general OR formula $e^{\sum \beta_i (X_{1i} - X_{0i})}$

When the model contains product terms, like CAT × AGE, or variables that are not (0, 1), like the continuous variable AGE, the simple formula will not work if the focus is on any of these variables. In such instances, we must use the general formula instead.

Chapters

✓ 1. Introduction

2. Important Special Cases

This presentation is now complete. We suggest that you review the material covered here by reading the summary section. You may also want to do the practice exercises and the test which follows. Then continue to the next chapter entitled, "Important Special Cases of the Logistic Model".

Detailed Outline

I. The multivariable problem (pages 4–5)

A. Example of a multivariate problem in epidemiologic research, including the issue of controlling for certain variables in the assessment of an exposure–disease relationship.

B. The general multivariate problem: assessment of the relationship of several independent variables, denoted as *X*s, to a dependent variable, denoted as *D*.

C. Flexibility in the types of independent variables allowed in most regression situations: A variety of variables are allowed.

D. Key restriction of model characteristics for the logistic model: The dependent variable is dichotomous.

II. Why is logistic regression popular? (pages 5–7)

A. Description of the logistic function.

B. Two key properties of the logistic function: Range is between 0 and 1 (good for describing probabilities) and the graph of function is S-shaped (good for describing combined risk factor effect on disease development).

III. The logistic model (pages 7–8)

A. Epidemiologic framework

B. Model formula:

$$\begin{aligned}\mathrm{P}(D=1|X_1,\ldots,X_k) &= \mathrm{P}(\mathbf{X})\\ &= 1/\{1+\exp[-(\alpha+\sum\beta_i X_i)]\}.\end{aligned}$$

IV. Applying the logistic model formula (pages 9–11)

A. The situation: independent variables CAT (0, 1), AGE (constant), ECG (0, 1); dependent variable CHD(0, 1); fit logistic model to data on 609 people.

B. Results for fitted model: estimated model parameters are

$\hat{\alpha} = -3.911, \hat{\beta}_1(\mathrm{CAT}) = 0.65, \hat{\beta}_2(\mathrm{AGE}) = 0.029,$ and $\hat{\beta}_3(\mathrm{ECG}) = 0.342.$

C. Predicted risk computations:

$\hat{\mathrm{P}}(\mathbf{X})$ for $\mathrm{CAT} = 1, \mathrm{AGE} = 40, \mathrm{ECG} = 0 : 0.1090,$

$\hat{\mathrm{P}}(\mathbf{X})$ for $\mathrm{CAT} = 0, \mathrm{AGE} = 40, \mathrm{ECG} = 0 : 0.0600.$

D. Estimated risk ratio calculation and interpretation: 0.1090/0.0600 = 1.82.

E. Risk ratio (RR) vs. odds ratio (OR): RR computation requires specifying all *X*s; OR is more natural measure for logistic model.

V. **Study design issues** (pages 11–15)

A. Follow-up orientation.

B. Applicability to case-control and cross-sectional studies? *Yes*.

C. Limitation in case-control and cross-sectional studies: cannot estimate risks, but can estimate odds ratios.

D. The limitation in mathematical terms: for case-control and cross-sectional studies, cannot get a good estimate of the constant.

VI. **Risk ratios vs. odds ratios** (pages 15–16)

A. Follow-up studies:

i. When all the variables in both groups compared are specified. [Example using CAT, AGE, and ECG comparing group 1 (CAT = 1, AGE = 40, ECG = 0) with group 0 (CAT = 0, AGE = 40, ECG = 0).]

ii. When control variables are unspecified, but assumed fixed and rare disease assumption is satisfied.

B. Case-control and cross-sectional studies: when rare disease assumption is satisfied.

C. What if rare disease assumption is not satisfied? Other approaches in the literature: Log-Binomial, Poisson, Copy method.

VII. **Logit transformation** (pages 16–22)

A. Definition of the logit transformation: logit $P(\mathbf{X}) = \ln_e[P(\mathbf{X})/(1 - P(\mathbf{X}))]$.

B. The formula for the logit function in terms of the parameters of the logistic model: logit $P(\mathbf{X}) = \alpha + \sum\beta_i X_i$.

C. Interpretation of the logit function in terms of odds:

i. $P(\mathbf{X})/[1 - P(\mathbf{X})]$ is the odds of getting the disease for an individual or group of individuals identified by **X**.

ii. The logit function describes the "log odds" for a person or group specified by **X**.

D. Interpretation of logistic model parameters in terms of log odds:

i. α is the log odds for a person or group when all *X*s are zero – can be critiqued on grounds that there is no such person.

ii. A more appealing interpretation is that α gives the "background or baseline" log odds, where "baseline" refers to a model that ignores all possible *X*s.

iii. The coefficient β_i represents the change in the log odds that would result from a one unit change in the variable X_{i} when all the other Xs are fixed.

iv. Example given for model involving CAT, AGE, and ECG: β_1 is the change in log odds corresponding to one unit change in CAT, when AGE and ECG are fixed.

VIII. Derivation of OR formula (pages 22–25)

A. Specifying two groups to be compared by an odds ratio: $\mathbf{X}_1$ and $\mathbf{X}_0$ denote the collection of Xs for groups 1 and 0.

B. Example involving CAT, AGE, and ECG variables: $\mathbf{X}_1 = (\mathrm{CAT} = 1, \mathrm{AGE} = 40, \mathrm{ECG} = 0)$, $\mathbf{X}_0 = (\mathrm{CAT} = 0, \mathrm{AGE} = 40, \mathrm{ECG} = 0)$.

C. Expressing the risk odds ratio (ROR) in terms of P($\mathbf{X}$):

$$\mathrm{ROR} = \frac{(\text{odds for } \mathbf{X}_1)}{(\text{odds for } \mathbf{X}_0)} = \frac{\mathrm{P}(\mathbf{X}_1)/1 - \mathrm{P}(\mathbf{X}_1)}{\mathrm{P}(\mathbf{X}_0)/1 - \mathrm{P}(\mathbf{X}_0)}.$$

D. Substitution of the model form for P($\mathbf{X}$) in the above ROR formula to obtain general ROR formula:

$$\mathrm{ROR} = \exp[\textstyle\sum \beta_i(X_{1i} - X_{0i})] = \Pi\{\exp[\beta_i(X_{1i} - X_{0i})]\}$$

E. Interpretation from the product (Π) formula: The contribution of each X_{i} variable to the odds ratio is *multiplicative*.

IX. Example of OR computation (pages 25–26)

A. Example of ROR formula for CAT, AGE, and ECG example using $\mathbf{X}_1$ and $\mathbf{X}_0$ specified in VIII B above: ROR = exp(β_1), where β_1 is the coefficient of CAT.

B. Interpretation of exp(β_1): an adjusted ROR for effect of CAT, controlling for AGE and ECG.

X. Special case for (0, 1) variables (pages 27–28)

A. General rule for (0, 1) variables: If variable is X_i, then ROR for effect of X_i controlling for other Xs in model is given by the formula ROR = exp(β_i), where β_i is the coefficient of X_i.

B. Example of formula in A for ECG, controlling for CAT and AGE.

C. Limitation of formula in A: Model can contain only main effect variables for Xs, and variable of focus must be (0, 1).

KEY FORMULAE

[$\exp(a) = e^{a}$ for any number a]

LOGISTIC FUNCTION: $f(z) = 1/[1 + \exp(-z)]$

LOGISTIC MODEL: $P(\mathbf{X}) = 1/\{1 + \exp[-(\alpha + \sum \beta_i X_i)]\}$

LOGIT TRANSFORMATION: logit $P(\mathbf{X}) = \alpha + \sum \beta_i X_i$

RISK ODDS RATIO (general formula):

$\text{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} := \exp[\sum \beta_i (X_{1i} - X_{0i})] = \Pi\{\exp[\beta_i (X_{1i} - X_{0i})]\}$

RISK ODDS RATIO [(0, 1) variables]: ROR $= \exp(\beta_i)$ for the effect of the variable X_i adjusted for the other Xs

Practice Exercises

Suppose you are interested in describing whether social status, as measured by a (0, 1) variable called SOC, is associated with cardiovascular disease mortality, as defined by a (0, 1) variable called CVD. Suppose further that you have carried out a 12-year follow-up study of 200 men who are 60 years old or older. In assessing the relationship between SOC and CVD, you decide that you want to control for smoking status [SMK, a (0, 1) variable] and systolic blood pressure (SBP, a continuous variable).

In analyzing your data, you decide to fit two logistic models, each involving the dependent variable CVD, but with different sets of independent variables. The variables involved in each model and their estimated coefficients are listed below:

Model 1		Model 2	
VARIABLE	COEFFICIENT	VARIABLE	COEFFICIENT
CONSTANT	−1.1800	CONSTANT	−1.1900
SOC	−0.5200	SOC	−0.5000
SBP	0.0400	SBP	0.0100
SMK	−0.5600	SMK	−0.4200
SOC × SBP	−0.0330		
SOC × SMK	0.1750		

1. For each of the models fitted above, state the form of the logistic model that was used (i.e., state the model in terms of the unknown population parameters and the independent variables being considered).

Model 1:

Model 2:

2. For each of the above models, state the form of the estimated model in logit terms.

   Model 1: logit P($\mathbf{X}$) =

   Model 2: logit P($\mathbf{X}$) =

3. Using Model 1, compute the estimated risk for CVD death (i.e., CVD = 1) for a high social class (SOC = 1) smoker (SMK = 1) with SBP = 150. (You will need a calculator to answer this. If you do not have one, just state the computational formula that is required, with appropriate variable values plugged in.)
4. Using Model 2, compute the estimated risk for CVD death for the following two persons:

   Person 1: SOC = 1, SMK = 1, SBP = 150.
   Person 2: SOC = 0, SMK = 1, SBP = 150.

   (As with the previous question, if you do not have a calculator, you may just state the computations that are required.)

   Person 1:

   Person 2:

5. Compare the estimated risk obtained in Exercise 3 with that for person 1 in Exercise 4. Why are not the two risks exactly the same?
6. Using Model 2 results, compute the risk ratio that compares person 1 with person 2. Interpret your answer.
7. If the study design had been either case-control or cross-sectional, could you have legitimately computed risk estimates as you did in the previous exercises? Explain.

8. If the study design had been case-control, what kind of measure of association could you have legitimately computed from the above models?
9. For Model 2, compute and interpret the estimated odds ratio for the effect of SOC, controlling for SMK and SBP? (Again, if you do not have a calculator, just state the computations that are required.)
10. Which of the following general formulae is *not* appropriate for computing the effect of SOC controlling for SMK and SBP in *Model 1*? (Circle one choice.) Explain your answer.
    a. $\exp(\beta_S)$, where β_S is the coefficient of SOC in model 1.
    b. $\exp[\sum \beta_i(X_{1i} - X_{0i})]$.
    c. $\Pi\{\exp[\beta_i(\mathrm{X}_{1i} - \mathrm{X}_{0i})]\}$.

Test

True or False (Circle T or F)

T F 1. We can use the logistic model provided all the independent variables in the model are continuous.

T F 2. Suppose the dependent variable for a certain multivariable analysis is systolic blood pressure, treated continuously. Then, a logistic model should be used to carry out the analysis.

T F 3. One reason for the popularity of the logistic model is that the range of the logistic function, from which the model is derived, lies between 0 and 1.

T F 4. Another reason for the popularity of the logistic model is that the shape of the logistic function is linear.

T F 5. The logistic model describes the probability of disease development, i.e., risk for the disease, for a given set of independent variables.

T F 6. The study design framework within which the logistic model is defined is a follow-up study.

T F 7. Given a fitted logistic model from case-control data, we can estimate the disease risk for a specific individual.

T F 8. In follow-up studies, we can use a fitted logistic model to estimate a risk ratio comparing two groups whenever all the independent variables in the model are specified for both groups.

T F 9. Given a fitted logistic model from a follow-up study, it is not possible to estimate individual risk as the constant term cannot be estimated.

T F 10. Given a fitted logistic model from a case-control study, an odds ratio can be estimated.

T F 11. Given a fitted logistic model from a case-control study, we can estimate a risk ratio if the rare disease assumption is appropriate.

T F 12. The logit transformation for the logistic model gives the log odds ratio for the comparison of two groups.

T F 13. The constant term, α, in the logistic model can be interpreted as a baseline log odds for getting the disease.

T F 14. The coefficient β_i in the logistic model can be interpreted as the change in log odds corresponding to a one unit change in the variable X_i that ignores the contribution of other variables.

T F 15. We can compute an odds ratio for a fitted logistic model by identifying two groups to be compared in terms of the independent variables in the fitted model.

T F 16. The product formula for the odds ratio tells us that the joint contribution of different independent variables to the odds ratio is additive.

T F 17. Given a (0, 1) independent variable and a model containing only main effect terms, the odds ratio that describes the effect of that variable controlling for the others in the model is given by e to the α, where α is the constant parameter in the model.

T F 18. Given independent variables AGE, SMK [smoking status (0, 1)], and RACE (0, 1), in a logistic model, an adjusted odds ratio for the effect of SMK is given by the natural log of the coefficient for the SMK variable.

T F 19. Given independent variables AGE, SMK, and RACE, as before, plus the product terms SMK × RACE and SMK × AGE, an adjusted odds ratio for the effect of SMK is obtained by exponentiating the coefficient of the SMK variable.

T F 20. Given the independent variables AGE, SMK, and RACE as in Question 18, but with SMK coded as (1, −1) instead of (0, 1), then e to the coefficient of the SMK variable gives the adjusted odds ratio for the effect of SMK.

21. Which of the following is *not* a property of the logistic model? (Circle one choice.)
    a. The model form can be written as $P(\mathbf{X})=1/\{1+\exp[-(\alpha+\sum\beta_i X_i)]\}$, where "exp{·}" denotes the quantity e raised to the power of the expression inside the brackets.
    b. logit $P(\mathbf{X}) = \alpha + \sum\beta_i X_i$ is an alternative way to state the model.
    c. $ROR = \exp[\sum\beta_i(X_{1i}-X_{0i})]$ is a general expression for the odds ratio that compares two groups of **X** variables.
    d. $ROR = \Pi\{\exp[\beta_i(X_{1i}-X_{0i})]\}$ is a general expression for the odds ratio that compares two groups of **X** variables.
    e. For any variable X_i, $ROR = \exp[\beta_i]$, where β_i is the coefficient of X_i, gives an adjusted odds ratio for the effect of X_i.

Suppose a logistic model involving the variables D = HPT [hypertension status (0, 1)], X_1 = AGE(continuous), X_2 = SMK(0, 1), X_3 = SEX(0, 1), X_4 = CHOL (cholesterol level, continuous), and X_5 = OCC[occupation (0, 1)] is fit to a set of data. Suppose further that the estimated coefficients of each of the variables in the model are given by the following table:

VARIABLE	COEFFICIENT
CONSTANT	−4.3200
AGE	0.0274
SMK	0.5859
SEX	1.1523
CHOL	0.0087
OCC	−0.5309

22. State the form of the logistic model that was fit to these data (i.e., state the model in terms of the unknown population parameters and the independent variables being considered).
23. State the form of the *estimated* logistic model obtained from fitting the model to the data set.
24. State the estimated logistic model in logit form.
25. Assuming the study design used was a follow-up design, compute the estimated risk for a 40-year-old male (SEX = 1) smoker (SMK = 1) with CHOL = 200 and OCC = 1. (You need a calculator to answer this question.)

26. Again assuming a follow-up study, compute the estimated risk for a 40-year-old male nonsmoker with CHOL = 200 and OCC = 1. (You need a calculator to answer this question.)
27. Compute and interpret the estimated risk ratio that compares the risk of a 40-year-old male smoker to a 40-year-old male nonsmoker, both of whom have CHOL = 200 and OCC = 1.
28. Would the risk ratio computation of Question 27 have been appropriate if the study design had been either cross-sectional or case-control? Explain.
29. Compute and interpret the estimated odds ratio for the effect of SMK controlling for AGE, SEX, CHOL, and OCC. (If you do not have a calculator, just state the computational formula required.)
30. What assumption will allow you to conclude that the estimate obtained in Question 29 is approximately a risk ratio estimate?
31. If you could not conclude that the odds ratio computed in Question 29 is approximately a risk ratio, what measure of association is appropriate? Explain briefly.
32. Compute and interpret the estimated odds ratio for the effect of OCC controlling for AGE, SMK, SEX, and CHOL. (If you do not have a calculator, just state the computational formula required.)
33. State two characteristics of the variables being considered in this example that allow you to use the $\exp(\beta_i)$ formula for estimating the effect of OCC controlling for AGE, SMK, SEX, and CHOL.
34. Why can you not use the formula $\exp(\beta_i)$ formula to obtain an adjusted odds ratio for the effect of AGE, controlling for the other four variables?

Answers to Practice Exercises

1. *Model* 1:
$$\hat{P}(\mathbf{X}) = 1/(1+\exp\{-[\alpha+\beta_1(\text{SOC})+\beta_2(\text{SBP}) + \beta_3(\text{SMK})+\beta_4(\text{SOC}\times\text{SBP}) + \beta_5(\text{SOC}\times\text{SMK})]\}).$$
Model 2:
$$\hat{P}(\mathbf{X}) = 1/(1+\exp\{-[\alpha+\beta_1(\text{SOC}) + \beta_2(\text{SBP})+\beta_3(\text{SMK})]\}).$$

2. *Model* 1:
$$\text{logit }\hat{\text{P}}(\mathbf{X}) = -1.18-0.52(\text{SOC})+0.04(\text{SBP}) - 0.56(\text{SMK})-0.033(\text{SOC}\times\text{SBP}) + 0.175(\text{SOC}\times\text{SMK}).$$
Model 2:
$$\text{logit }\hat{\text{P}}(\mathbf{X}) = -1.19-0.50(\text{SOC})+0.01(\text{SBP}) - 0.42(\text{SMK}).$$

3. For SOC = 1, SBP = 150, and SMK = 1, **X** = (SOC, SBP, SMK, SOC×SBP, SOC×SMK) = (1, 150, 1, 150, 1) and

$$\begin{aligned} \textit{Model }1, \hat{\mathrm{P}}(\mathbf{X}) &= 1/(1+\exp\{-[-1.18-0.52(1) \\ &\quad +0.04(150)-0.56(1) \\ &\quad -0.033(1\times 150)-0.175(1\times 1)]\}). \\ &= 1/\{1+\exp[-(-1.035)]\} \\ &= 1/(1+2.815) \\ &= 0.262 \end{aligned}$$

4. For *Model 2, person 1* (SOC = 1, SMK = 1, SBP = 150):

$$\begin{aligned} \hat{\mathrm{P}}(\mathbf{X}) &= 1/(1+\exp\{-[-1.19-0.50(1) \\ &\quad +0.01(150)-0.42(1)]\}) \\ &= 1/\{1+\exp[-(-0.61)]\} \\ &= 1/(1+1.84) \\ &= 0.352 \end{aligned}$$

For *Model 2, person 2* (SOC = 0, SMK = 1, SBP = 150):

$$\begin{aligned} \hat{\mathrm{P}}(\mathbf{X}) &= 1/(1+\exp\{-[-1.19-0.50(0) \\ &\quad +0.01(150)-0.42(1)]\}) \\ &= 1/\{1+\exp[-(-0.11)]\} \\ &= 1/(1+1.116) \\ &= 0.473 \end{aligned}$$

5. The risk computed for *Model 1* is 0.262, whereas the risk computed for *Model 2, person 1* is 0.352. Note that both risks are computed for the same person (i.e., SOC = 1, SMK = 1, SBP = 150), yet they yield different values because the models are different. In particular, *Model 1* contains two product terms that are not contained in *Model 2*, and consequently, computed risks for a given person can be expected to be somewhat different for different models.

6. Using *Model 2* results,

$$\begin{aligned} \mathrm{RR}(1\text{ vs. }2) &= \frac{\mathrm{P}(\mathrm{SOC}=0, \mathrm{SMK}=1, \mathrm{SBP}=150)}{\mathrm{P}(\mathrm{SOC}=1, \mathrm{SMK}=1, \mathrm{SBP}=150)} \\ &= 0.352/0.473 = 1/1.34 = 0.744 \end{aligned}$$

This estimated risk ratio is less than 1 because the risk for high social class persons (SOC = 1) is less than the risk for low social class persons (SOC = 0) in this data set. More specifically, the risk for low social class persons is 1.34 times as large as the risk for high social class persons.

7. No. If the study design had been either case-control or cross-sectional, risk estimates could not be computed because the constant term (α) in the model could not be estimated. In other words, even if the computer printed out values of -1.18 or -1.19 for the constant terms, these numbers would not be legitimate estimates of α.
8. For case-control studies, only odds ratios, not risks or risk ratios, can be computed directly from the fitted model.
9. $\widehat{\text{OR}}(\text{SOC} = 1 \text{ vs. SOC} = 0 \text{ controlling for SMK and SBP})$

   $= e^{\hat{\beta}}$, where $\hat{\beta} = -0.50$ is the estimated coefficient of SOC in the fitted model

   $= \exp(-0.50)$

   $= 0.6065 = 1/1.65.$

   The estimated odds ratio is less than 1, indicating that, for this data set, the risk of CVD death for high social class persons is less than the risk for low social class persons. In particular, the risk for low social class persons is estimated as 1.65 times as large as the risk for high social class persons.
10. Choice (a) is *not* appropriate for the effect of SOC using model 1. Model 1 contains interaction terms, whereas choice (a) is appropriate only if all the variables in the model are main effect terms. Choices (b) and (c) are two equivalent ways of stating the general formula for calculating the odds ratio for any kind of logistic model, regardless of the types of variables in the model.

2 Important Special Cases of the Logistic Model

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_2, © Springer Science+Business Media, LLC 2010

Introduction

In this chapter, several important special cases of the logistic model involving a single (0, 1) exposure variable are considered with their corresponding odds ratio expressions. In particular, focus is on defining the independent variables that go into the model and on computing the odds ratio for each special case. Models that account for the potential confounding effects and potential interaction effects of covariates are emphasized.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

Objectives

Upon completion of this chapter, the learner should be able to:

1. State or recognize the logistic model for a simple analysis.
2. Given a model for simple analysis:
   a. state an expression for the odds ratio describing the exposure–disease relationship
   b. state or recognize the null hypothesis of no exposure–disease relationship in terms of parameter(s) of the model
   c. compute or recognize an expression for the risk for exposed or unexposed persons separately
   d. compute or recognize an expression for the odds of getting the disease for exposed or unexposed persons separately
3. Given two (0, 1) independent variables:
   a. state or recognize a logistic model that allows for the assessment of interaction on a multiplicative scale
   b. state or recognize the expression for no interaction on a multiplicative scale in terms of odds ratios for different combinations of the levels of two (0, 1) independent variables
   c. state or recognize the null hypothesis for no interaction on a multiplicative scale in terms of one or more parameters in an appropriate logistic model
4. Given a study situation involving a (0, 1) exposure variable and several control variables:
   a. state or recognize a logistic model that allows for the assessment of the exposure-disease relationship, controlling for the potential confounding and potential interaction effects of functions of the control variables
   b. compute or recognize the expression for the odds ratio for the effect of exposure on disease status adjusting for the potential confounding and interaction effects of the control variables in the model
   c. state or recognize an expression for the null hypothesis of no interaction effect involving one or more of the effect modifiers in the model
   d. assuming no interaction, state or recognize an expression for the odds ratio for the effect of exposure on disease status adjusted for confounders

e. assuming no interaction, state or recognize the null hypothesis for testing the significance of this odds ratio in terms of a parameter in the model

5. Given a logistic model involving interaction terms, state or recognize that the expression for the odds ratio will give different values for the odds ratio depending on the values specified for the effect modifiers in the model.

Presentation

I. Overview

Special Cases:

- Simple analysis $\left(\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}\right)$
- Multiplicative interaction
- Controlling several confounders and effect modifiers

This presentation describes important special cases of the general logistic model when there is a single (0, 1) exposure variable. Special case models include simple analysis of a fourfold table, assessment of multiplicative interaction between two dichotomous variables, and controlling for several confounders and interaction terms. In each case, we consider the definitions of variables in the model and the formula for the odds ratio describing the exposure-disease relationship.

General logistic model formula:

$$\mathrm{P}(\mathbf{X}) = \frac{1}{1 + \mathrm{e}^{-\left(\alpha + \sum \beta_i \mathrm{X}_i\right)}}$$

$\mathbf{X} = (X_1, X_2, \ldots, X_k)$

α, β_i = unknown parameters

D = dichotomous outcome

Recall that the general logistic model for k independent variables may be written as P(**X**) equals 1 over 1 plus e to minus the quantity α plus the sum of $\beta_i X_i$, where P(**X**) denotes the probability of developing a disease of interest given values of a collection of independent variables X_1, X_2, through X_k, that are collectively denoted by the *bold* **X**. The terms α and β_i in the model represent unknown parameters that we need to estimate from data obtained for a group of subjects on the Xs and on D, a dichotomous disease outcome variable.

$$\operatorname{logit} \mathrm{P}(\mathbf{X}) = \underbrace{\alpha + \textstyle\sum \beta_i \mathrm{X}_i}_{\text{linear sum}}$$

An alternative way of writing the logistic model is called the logit form of the model. The expression for the logit form is given here.

$$\mathrm{ROR} = \mathrm{e}^{\sum\limits_{i=1}^{k} \beta_i (X_{1i} - X_{0i})}$$

$$= \prod_{i=1}^{k} \mathrm{e}^{\beta_i (X_{1i} - X_{0i})}$$

$\mathbf{X}_1$	specification of **X** for subject 1
$\mathbf{X}_0$	specification of **X** for subject 0

The general odds ratio formula for the logistic model is given by either of two formulae. The first formula is of the form e to a sum of linear terms. The second is of the form of the product of several exponentials; that is, each term in the product is of the form e to some power. Either formula requires two specifications, $\mathbf{X}_1$ and $\mathbf{X}_0$, of the collection of k independent variables X_1, $X_2, \ldots, X_k$.

We now consider a number of important special cases of the logistic model and their corresponding odds ratio formulae.

II. Special Case – Simple Analysis

$X_1 = E =$ exposure (0, 1)

$D =$ disease (0, 1)

We begin with the simple situation involving one dichotomous independent variable, which we will refer to as an *exposure* variable and will denote it as $X_1 = E$. Because the disease variable, D, considered by a logistic model is dichotomous, we can use a two-way table with four cells to characterize this analysis situation, which is often referred to as a *simple analysis*.

	$E = 1$	$E = 0$
$D = 1$	a	b
$D = 0$	c	d

For convenience, we define the exposure variable as a (0, 1) variable and place its values in the two columns of the table. We also define the disease variable as a (0, 1) variable and place its values in the rows of the table. The cell frequencies within the fourfold table are denoted as a, b, c, and d, as is typically presented for such a table.

$$\mathrm{P}(\mathbf{X}) = \frac{1}{1 + \mathrm{e}^{-(\alpha+\beta_1 E)}},$$

where $E = (0, 1)$ variable.

Note: Other coding schemes (1, −1), (1, 2), (2, 1)

A logistic model for this simple analysis situation can be defined by the expression P(**X**) equals 1 over 1 plus e to minus the quantity α plus β_1 times E, where E takes on the value 1 for exposed persons and 0 for unexposed persons. Note that other coding schemes for E are also possible, such as (1, −1), (1, 2), or even (2, 1). However, we defer discussing such alternatives until Chap. 3.

$$\text{logit } \mathrm{P}(\mathbf{X}) = \alpha + \beta_1 E$$

The logit form of the logistic model we have just defined is of the form logit P(**X**) equals the simple linear sum α plus β_1 times E. As stated earlier in our review, this logit form is an alternative way to write the statement of the model we are using.

$\mathrm{P}(\mathbf{X}) = \Pr(D = 1|E)$

$E = 1$: $\mathrm{R}_1 = \Pr(D = 1|E = 1)$

$E = 0$: $\mathrm{R}_0 = \Pr(D = 1|E = 0)$

The term P(**X**) for the simple analysis model denotes the probability that the disease variable D takes on the value 1, given whatever the value is for the exposure variable E. In epidemiologic terms, this probability denotes the *risk* for developing the disease, given exposure status. When the value of the exposure variable equals 1, we call this risk $\mathbf{R}_1$, which is the conditional probability that D equals 1 given that E equals 1. When E equals 0, we denote the risk by $\mathbf{R}_0$, which is the conditional probability that D equals 1 given that E equals 0.

$$\text{ROR}_{E=1\text{ vs. }E=0} = \frac{\dfrac{\mathbf{R}_1}{1-\mathbf{R}_1}}{\dfrac{\mathbf{R}_0}{1-\mathbf{R}_0}}$$

We would like to use the above model for simple analysis to obtain an expression for the odds ratio that compares exposed persons with unexposed persons. Using the terms $\mathbf{R}_1$ and $\mathbf{R}_0$, we can write this odds ratio as $\mathbf{R}_1$ divided by 1 minus $\mathbf{R}_1$ over $\mathbf{R}_0$ divided by 1 minus $\mathbf{R}_0$.

Substitute $\text{P}(\mathbf{X}) = \dfrac{1}{1 + e^{-\left(\alpha+\sum\beta_i \text{X}_i\right)}}$ into ROR formula:

To compute the odds ratio in terms of the parameters of the logistic model, we substitute the logistic model expression into the odds ratio formula.

$$E = 1: \ \mathbf{R}_1 = \frac{1}{1+e^{-(\alpha+[\beta_1\times 1])}} = \frac{1}{1+e^{-(\alpha+\beta_1)}}$$

For E equal to 1, we can write $\mathbf{R}_1$ by substituting the value E equals 1 into the model formula for $\text{P}(\mathbf{X})$. We then obtain 1 over 1 plus e to minus the quantity α plus β_1 times 1, or simply 1 over 1 plus e to minus α plus β_1.

$$E = 0: \ \mathbf{R}_0 = \frac{1}{1+e^{-(\alpha+[\beta_1\times 0])}} = \frac{1}{1+e^{-\alpha}}$$

For E equal to zero, we write $\mathbf{R}_0$ by substituting E equal to 0 into the model formula, and we obtain 1 over 1 plus e to minus α.

$$\text{ROR} = \frac{\dfrac{\mathbf{R}_1}{1-\mathbf{R}_1}}{\dfrac{\mathbf{R}_0}{1-\mathbf{R}_0}} = \frac{\dfrac{1}{1+e^{-(\alpha+\beta_1)}}}{\dfrac{1}{1+e^{-\alpha}}}$$

algebra
$$= \boxed{e^{\beta_1}}$$

To obtain ROR then, we replace $\mathbf{R}_1$ with 1 over 1 plus e to minus α plus β_1, and we replace $\mathbf{R}_0$ with 1 over 1 plus e to minus α. The ROR formula then simplifies algebraically to e to the β_1, where β_1 is the coefficient of the exposure variable.

General ROR formula used for other special cases

We could have obtained this expression for the odds ratio using the general formula for the ROR that we gave during our review. We will use the general formula now. Also, for other special cases of the logistic model, we will use the general formula rather than derive an odds ratio expression separately for each case.

General:

$$\text{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} = e^{\sum\limits_{i=1}^{k}\beta_i(X_{1i}-X_{0i})}$$

Simple analysis:

$k = 1, \mathbf{X} = (X_1), \beta_i = \beta_1$

The general formula computes ROR as e to the sum of each β_i times the difference between X_{1i} and X_{0i}, where X_{1i} denotes the value of the ith X variable for group 1 persons and X_{0i} denotes the value of the ith X variable for group 0 persons. In a simple analysis, we have only one X and one β; in other words, k, the number of variables in the model, equals 1.

group 1: $\mathbf{X}_1 = E = 1$
group 0: $\mathbf{X}_0 = E = 0$

$\mathbf{X}_1 = (X_{11}) = (1)$
$\mathbf{X}_0 = (X_{01}) = (0)$

For a simple analysis model, group 1 corresponds to exposed persons, for whom the variable X_1, in this case E, equals 1. Group 0 corresponds to unexposed persons, for whom the variable X_1 or E equals 0. Stated another way, for group 1, the collection of Xs denoted by the *bold* $\mathbf{X}$ can be written as $\mathbf{X}_1$ and equals the collection of one value X_{11}, which equals 1. For group 0, the collection of Xs denoted by the *bold* $\mathbf{X}$ is written as $\mathbf{X}_0$ and equals the collection of one value X_{01}, which equals 0.

$$\begin{aligned}\text{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} &= e^{\beta_1(X_{11}-X_{01})}\\ &= e^{\beta_1(1-0)}\\ &= e^{\beta_1}\end{aligned}$$

Substituting the particular values of the one X variable into the general odds ratio formula then gives e to the β_1 times the quantity X_{11} minus X_{01}, which becomes e to the β_1 times 1 minus 0, which reduces to e to the β_1.

SIMPLE ANALYSIS SUMMARY

$$P(\mathbf{X}) = \frac{1}{1 + e^{-(\alpha+\beta_1 E)}}$$

$$\text{ROR} = e^{\beta_1}$$

In summary, for the simple analysis model involving a (0, 1) exposure variable, the logistic model P($\mathbf{X}$) equals 1 over 1 plus e to minus the quantity α plus β_1 times E, and the odds ratio that describes the effect of the exposure variable is given by e to the β_1, where β_1 is the coefficient of the exposure variable.

$$\widehat{\text{ROR}}_{\mathbf{X}_1,\,\mathbf{X}_0} = e^{\hat{\beta}_1}$$

We can estimate this odds ratio by fitting the simple analysis model to a set of data. The estimate of the parameter β_1 is typically denoted as $\hat{\beta}_1$. The odds ratio estimate then becomes e to the $\hat{\beta}_1$.

	$E = 1$	$E = 0$
$D = 1$	a	b
$D = 0$	c	d

$\widehat{\text{ROR}} = e^{\hat{\beta}} = ad/bc$

The reader should not be surprised to find out that an alternative formula for the estimated odds ratio for the simple analysis model is the familiar *a* times *d* over *b* times *c*, where *a, b, c,* and *d* are the cell frequencies in the fourfold table for simple analysis. That is, e to the $\hat{\beta}_1$ obtained from fitting a logistic model for simple analysis can alternatively be computed as *ad* divided by *bc* from the cell frequencies of the fourfold table.

Simple analysis: does not need computer

Other special cases: require computer

Thus, in the simple analysis case, we need not go to the trouble of fitting a logistic model to get an odds ratio estimate as the typical formula can be computed without a computer program. We have presented the logistic model version of simple analysis to show that the logistic model incorporates simple analysis as a special case. More complicated special cases, involving more than one independent variable, require a computer program to compute the odds ratio.

III. Assessing Multiplicative Interaction

We will now consider how the logistic model allows the assessment of interaction between two independent variables.

$X_1 = A = (0, 1)$ variable

$X_2 = B = (0, 1)$ variable

Interaction: equation involving RORs for combinations of A and B

Consider, for example, two (0, 1) *X* variables, X_1 and X_2, which for convenience we rename as *A* and *B*, respectively. We first describe what we mean conceptually by interaction between these two variables. This involves an equation involving risk odds ratios corresponding to different combinations of *A* and *B*. The odds ratios are defined in terms of risks, which we now describe.

$$\begin{aligned} R_{AB} &= \text{risk given } A, B \\ &= \Pr(D = 1 \mid A, B) \end{aligned}$$

Let R_{AB} denote the risk for developing the disease, given specified values for *A* and *B*; in other words, R_{AB} equals the conditional probability that *D* equals 1, given *A* and *B*.

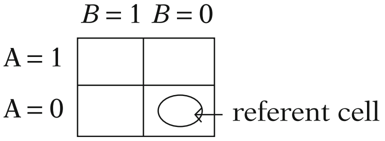
	$B = 1$	$B = 0$
A = 1	R_{11}	R_{10}
A = 0	R_{01}	R_{00}

Because *A* and *B* are dichotomous, there are four possible values for R_{AB}, which are shown in the cells of a two-way table. When *A* equals 1 and *B* equals 1, the risk R_{AB} becomes R_{11}. Similarly, when *A* equals 1 and B equals 0, the risk becomes R_{10}. When *A* equals 0 and B equals 1, the risk is R_{01}, and finally, when *A* equals 0 and B equals 0, the risk is R_{00}.

Note: above table not for simple analysis.

	$B = 1$	$B = 0$
$A = 1$	R_{11}	R_{10}
$A = 0$	R_{01}	R_{00}

Note that the two-way table presented here does not describe a simple analysis because the row and column headings of the table denote two independent variables rather than one independent variable and one disease variable. Moreover, the information provided within the table is a collection of four risks corresponding to different combinations of both independent variables, rather than four cell frequencies corresponding to different exposure-disease combinations.



$\text{OR}_{11} = \text{odds}(1, 1)/\text{odds}(0, 0)$

$\text{OR}_{10} = \text{odds}(1, 0)/\text{odds}(0, 0)$

$\text{OR}_{01} = \text{odds}(0, 1)/\text{odds}(0, 0)$

Within this framework, odds ratios can be defined to compare the odds for any one cell in the two-way table of risks with the odds for any other cell. In particular, three odds ratios of typical interest compare each of three of the cells to a *referent cell*. The referent cell is usually selected to be the combination A equals 0 and B equals 0. The three odds ratios are then defined as OR_{11}, OR_{10}, and OR_{01}, where OR_{11} equals the odds for cell 11 divided by the odds for cell 00, OR_{10} equals the odds for cell 10 divided by the odds for cell 00, and OR_{01} equals the odds for cell 01 divided by the odds for cell 00.

odds $(A,B) = R_{AB}/(1 - R_{AB})$

$$\text{OR}_{11} = \frac{R_{11}/(1-R_{11})}{R_{00}/(1-R_{00})} = \frac{R_{11}(1-R_{00})}{R_{00}(1-R_{11})}$$

$$\text{OR}_{10} = \frac{R_{10}/(1-R_{10})}{R_{00}/(1-R_{00})} = \frac{R_{10}(1-R_{00})}{R_{00}(1-R_{10})}$$

$$\text{OR}_{01} = \frac{R_{01}/(1-R_{01})}{R_{00}/(1-R_{00})} = \frac{R_{01}(1-R_{00})}{R_{00}(1-R_{01})}$$

As the odds for any cell A,B is defined in terms of risks as R_{AB} divided by 1 minus R_{AB}, we can obtain the following expressions for the three odds ratios: OR_{11} equals the product of R_{11} times 1 minus R_{00} divided by the product of R_{00} times 1 minus R_{11}. The corresponding expressions for OR_{10} and OR_{01} are similar, where the subscript 11 in the numerator and denominator of the 11 formula is replaced by 10 and 01, respectively.

$$\text{OR}_{\text{AB}} = \frac{R_{\text{AB}}(1-R_{00})}{R_{00}(1-R_{\text{AB}})}$$

$A = 0, 1; \quad B = 0, 1$

In general, without specifying the value of A and B, we can write the odds ratio formulae as OR_{AB} equals the product of R_{AB} and 1 minus R_{00} divided by the product of R_{00} and $1 - R_{AB}$, where A takes on the values 0 and 1 and B takes on the values 0 and 1.

DEFINITION

$$\mathrm{OR}_{11} = \mathrm{OR}_{10} \underset{\text{multiplication}}{\times} \mathrm{OR}_{01}$$

no interaction on a multiplicative scale

Now that we have defined appropriate odds ratios for the two independent variables situation, we are ready to provide an equation for assessing interaction. The equation is stated as OR_{11} equals the product of OR_{10} and OR_{01}. If this expression is satisfied for a given study situation, we say that there is "no interaction on a *multiplicative* scale." In contrast, if this expression is not satisfied, we say that there is evidence of interaction on a multiplicative scale.

Note that the right-hand side of the "no interaction" expression requires *multiplication* of two odds ratios, one corresponding to the combination 10 and the other to the combination 01. Thus, the scale used for assessment of interaction is called multiplicative.

No interaction:

$$\underset{\mathrm{OR}_{11}}{\underset{\uparrow}{\begin{pmatrix}\text{effect of}\\ A \text{ and } B\\ \text{acting}\\ \text{together}\end{pmatrix}}} = \underset{\substack{\mathrm{OR}_{10}\times\mathrm{OR}_{01}\\ \text{multiplicative}\\ \text{scale}}}{\underset{\uparrow}{\begin{pmatrix}\text{combined}\\ \text{effect of}\\ \text{A and B}\\ \text{acting}\\ \text{separately}\end{pmatrix}}}$$

When the no interaction equation is satisfied, we can interpret the effect of both variables *A* and *B* acting together as being the same as the combined effect of each variable acting separately.

The effect of both variables acting together is given by the odds ratio OR_{11} obtained when *A* and *B* are both present, that is, when *A* equals 1 and *B* equals 1.

The effect of *A* acting separately is given by the odds ratio for *A* equals 1 and *B* equals 0, and the effect of *B* acting separately is given by the odds ratio for *A* equals 0 and *B* equals 1. The combined separate effects of *A* and *B* are then given by the product OR_{10} times OR_{01}.

no interaction formula:
$\mathrm{OR}_{11} = \mathrm{OR}_{10} \times \mathrm{OR}_{01}$

Thus, when there is no interaction on a multiplicative scale, OR_{11} equals the product of OR_{10} and OR_{01}.

EXAMPLE

	$B = 1$	$B = 0$
$A = 1$	$R_{11} = 0.0350$	$R_{10} = 0.0175$
$A = 0$	$R_{01} = 0.0050$	$R_{00} = 0.0025$

$$\mathrm{OR}_{11} = \frac{0.0350(1 - 0.0025)}{0.0025(1 - 0.0350)} = 14.4$$

$$\mathrm{OR}_{10} = \frac{0.0175(1 - 0.0025)}{0.0025(1 - 0.0175)} = 7.2$$

$$\mathrm{OR}_{01} = \frac{0.0050(1 - 0.0025)}{0.0025(1 - 0.0050)} = 2.0$$

$$\mathrm{OR}_{11} \overset{?}{=} \mathrm{OR}_{10} \times \mathrm{OR}_{01}$$

$$14.4 \underset{\substack{\uparrow \\ \boxed{\text{Yes}}}}{\overset{?}{=}} \underbrace{7.2 \times 2.0}_{14.4}$$

$B = 1$	$B = 0$
$R_{11} = 0.0700$	$R_{10} = 0.0175$
$R_{01} = 0.0050$	$R_{00} = 0.0025$

$\mathrm{OR}_{11} = 30.0$

$\mathrm{OR}_{10} = 7.2$

$\mathrm{OR}_{01} = 2.0$

$$\mathrm{OR}_{11} \overset{?}{=} \mathrm{OR}_{10} \times \mathrm{OR}_{01}$$

$$30.0 \underset{\substack{\uparrow \\ \boxed{\text{No}}}}{\overset{?}{=}} 7.2 \times 2.0$$

As an example of no interaction on a multiplicative scale, suppose the risks R_{AB} in the fourfold table are given by R_{11} equal to 0.0350, R_{10} equal to 0.0175, R_{01} equal to 0.0050, and R_{00} equal to 0.0025. Then the corresponding three odds ratios are obtained as follows: OR_{11} equals 0.0350 times 1 minus 0.0025 divided by the product of 0.0025 and 1 minus 0.0350, which becomes 14.4; OR_{10} equals 0.0175 times 1 minus 0.0025 divided by the product of 0.0025 and 1 minus 0.0175, which becomes 7.2; and OR_{01} equals 0.0050 times 1 minus 0.0025 divided by the product of 0.0025 and 1 minus 0.0050, which becomes 2.0.

To see if the no interaction equation is satisfied, we check whether OR_{11} equals the product of OR_{10} and OR_{01}. Here we find that OR_{11} equals 14.4 and the product of OR_{10} and OR_{01} is 7.2 times 2, which is also 14.4. Thus, the no interaction equation is satisfied.

In contrast, using a different example, if the risk for the 11 cell is 0.0700, whereas the other three risks remained at 0.0175, 0.0050, and 0.0025, then the corresponding three odds ratios become OR_{11} equals 30.0, OR_{10} equals 7.2, and OR_{01} equals 2.0. In this case, the no interaction equation is not satisfied because the left-hand side equals 30 and the product of the two odds ratios on the right-hand side equals 14. Here, then, we would conclude that there is interaction because the effect of both variables acting together is more than twice the combined effect of the variables acting separately.

EXAMPLE (continued)

Note: "=" means approximately equal ($\approx$)
e.g., $14.5 \approx 14.0 \Rightarrow$ no interaction

Note that in determining whether or not the no interaction equation is satisfied, the left- and right-hand sides of the equation do not have to be exactly equal. If the left-hand side is approximately equal to the right-hand side, we can conclude that there is no interaction. For instance, if the left-hand side is 14.5 and the right-hand side is 14, this would typically be close enough to conclude that there is no interaction on a multiplicative scale.

REFERENCE
multiplicative interaction vs. additive interaction
Epidemiologic Research, Chap. 19

A more complete discussion of interaction, including the distinction between *multiplicative interaction* and *additive interaction*, is given in Chap. 19 of *Epidemiologic Research* by Kleinbaum, Kupper, and Morgenstern (1982).

Logistic model variables:

$$\left.\begin{aligned} X_1 &= A_{(0,1)} \\ X_2 &= B_{(0,1)} \end{aligned}\right\} \text{main effects}$$

$X_3 = A \times B$ interaction effect variable

We now define a logistic model that allows the assessment of multiplicative interaction involving two (0, 1) indicator variables A and B. This model contains three independent variables, namely, X_1 equal to A, X_2 equal to B, and X_3 equal to the product term A times B. The variables A and B are called main effect variables and the product term is called an interaction effect variable.

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1 A + \beta_2 B + \beta_3 A \times B,$$

where

$$\begin{aligned} \text{P(X)} &= \text{risk given } A \text{ and } B \\ &= R_{AB} \end{aligned}$$

The logit form of the model is given by the expression logit of P(**X**) equals α plus β_1 times A plus β_2 times B plus β_3 times A times B. P(**X**) denotes the risk for developing the disease given values of A and B, so that we can alternatively write P(**X**) as R_{AB}.

$$\beta_3 = \ln_e \left[\frac{\text{OR}_{11}}{\text{OR}_{10} \times \text{OR}_{01}} \right]$$

For this model, it can be shown mathematically that the coefficient β_3 of the product term can be written in terms of the three odds ratios we have previously defined. The formula is β_3 equals the natural log of the quantity OR_{11} divided by the product of OR_{10} and OR_{01}. We can make use of this formula to test the null hypothesis of no interaction on a multiplicative scale.

H_0 no interaction on a multiplicative scale

$$\Leftrightarrow H_0: \mathrm{OR}_{11} = \mathrm{OR}_{10} \times \mathrm{OR}_{01}$$

$$\Leftrightarrow H_0: \frac{\mathrm{OR}_{11}}{\mathrm{OR}_{10} \times \mathrm{OR}_{01}} = 1$$

$$\Leftrightarrow H_0: \ln_e\left(\frac{\mathrm{OR}_{11}}{\mathrm{OR}_{10} \times \mathrm{OR}_{01}}\right) = \ln_e 1$$

$$\Leftrightarrow H_0: \beta_3 = 0$$

One way to state this null hypothesis, as described earlier in terms of odds ratios, is OR_{11} equals the product of OR_{10} and OR_{01}. Now it follows algebraically that this odds ratio expression is equivalent to saying that the quantity OR_{11} divided by OR_{10} times OR_{01} equals 1, or equivalently, that the natural log of this expression equals the natural log of 1, or, equivalently, that β_3 equals 0. Thus, the null hypothesis of no interaction on a multiplicative scale can be equivalently stated as β_3 equals 0.

$$\text{logit } P(\mathbf{X}) = \alpha + \beta_1 A + \beta_2 B + \beta_3 AB$$

H_0: no interaction $\Leftrightarrow \beta_3 = 0$

Test result	*Model*
not significant	$\Rightarrow \alpha + \beta_1 A + \beta_2 B$
significant	$\Rightarrow \alpha + \beta_1 A + \beta_2 B + \beta_3 AB$

In other words, a test for the no interaction hypotheses can be obtained by testing for the significance of the coefficient of the product term in the model. If the test is not significant, we would conclude that there is no interaction on a multiplicative scale and we would reduce the model to a simpler one involving only main effects. In other words, the reduced model would be of the form logit $P(\mathbf{X})$ equals α plus β_1 times A plus β_2 times B. If, on the other hand, the test is significant, the model would retain the β_3 term and we would conclude that there is significant interaction on a multiplicative scale.

MAIN POINT:
Interaction test ⇒ test for product terms

A description of methods for testing hypotheses for logistic regression models is beyond the scope of this presentation (see Chap. 5). The main point here is that we can test for interaction in a logistic model by testing for significance of product terms that reflect interaction effects in the model.

EXAMPLE

Case-control study

ASB = (0, 1)	variable for asbestos exposure
SMK = (0, 1)	variable for smoking status
D = (0, 1)	variable for bladder cancer status

As an example of a test for interaction, we consider a study that looks at the combined relationship of asbestos exposure and smoking to the development of bladder cancer. Suppose we have collected case-control data on several persons with the same occupation. We let *ASB* denote a (0,1) variable indicating asbestos exposure status, *SMK* denote a (0, 1) variable indicating smoking status, and *D* denote a (0, 1) variable for bladder cancer status.

EXAMPLE (continued)

$$\text{logit}\ (\mathbf{X}) = \alpha + \beta_1\text{ASB} + \beta_2\text{SMK} + \beta_3\text{ASB} \times \text{SMK}$$

$\mathbf{H}_0$: no interaction (multiplicative)
$\Leftrightarrow \text{H}_0 : \beta_3 = 0$

Test Result	*Conclusion*
Not Significant	No interaction on multiplicative scale
Significant ($\hat{\beta}_3 > 0$)	Joint effect > combined effect
Significant ($\hat{\beta}_3 < 0$)	Joint effect < combined effect

To assess the extent to which there is a multiplicative interaction between asbestos exposure and smoking, we consider a logistic model with ASB and SMK as main effect variables and the product term ASB times SMK as an interaction effect variable. The model is given by the expression logit P($\mathbf{X}$) equals α plus β_1 times ASB plus β_2 times SMK plus β_3 times ASB times SMK. With this model, a test for no interaction on a multiplicative scale is equivalent to testing the null hypothesis that β_3, the coefficient of the product term, equals 0.

If this test is not significant, then we would conclude that the effect of asbestos and smoking acting together is equal, on a multiplicative scale, to the combined effect of asbestos and smoking acting separately. If this test is significant and $\hat{\beta}_3$ is greater than 0, we would conclude that the joint effect of asbestos and smoking is greater than a multiplicative combination of separate effects. Or, if the test is significant and $\hat{\beta}_3$ is less than zero, we would conclude that the joint effect of asbestos and smoking is less than a multiplicative combination of separate effects.

IV. The *E, V, W* Model – A General Model Containing a (0, 1) Exposure and Potential Confounders and Effect Modifiers

The variables:
E = (0, 1) exposure
$C_1, C_2, \ldots, C_p$ continuous or categorical

We are now ready to discuss a logistic model that considers the effects of several independent variables and, in particular, allows for the control of confounding and the assessment of interaction. We call this model the *E, V, W* model. We consider a single dichotomous (0, 1) exposure variable, denoted by E, and p extraneous variables C_1, C_2, and so on, up through C_p. The variables C_1 through C_p may be either continuous or categorical.

EXAMPLE

$$\begin{aligned} D &= \text{CHD}_{(0,1)} \\ E &= \text{CAT}_{(0,1)} \end{aligned}$$

$$\text{Control variables} \begin{cases} C_1 = \text{AGE}_{\text{continous}} \\ C_2 = \text{CHL}_{\text{continous}} \\ C_3 = \text{SMK}_{(0,1)} \\ C_4 = \text{ECG}_{(0,\,1)} \\ C_5 = \text{HPT}_{(0,\,1)} \end{cases}$$

As an example of this special case, suppose the disease variable is coronary heart disease status (CHD), the exposure variable E is catecholamine level (CAT), where 1 equals high and 0 equals low, and the control variables are AGE, cholesterol level (CHL), smoking status (SMK), electrocardiogram abnormality status (ECG), and hypertension status (HPT).

EXAMPLE (continued)

1 E : CAT
5 *C*s : AGE, CHL, SMK, ECG, HPT

We will assume here that both AGE and CHL are treated as continuous variables, that SMK is a (0, 1) variable, where 1 equals ever smoked and 0 equals never smoked, that ECG is a (0, 1) variable, where 1 equals abnormality present and 0 equals abnormality absent, and that HPT is a (0, 1) variable, where 1 equals high blood pressure and 0 equals normal blood pressure. There are, thus, five *C* variables in addition to the exposure variable CAT.

Model with eight independent variables:

2 E × *C*s : CAT × CHL
CAT × HPT

We now consider a model with eight independent variables. In addition to the exposure variable CAT, the model contains the five *C* variables as potential confounders plus two product terms involving two of the *C*s, namely, CHL and HPT, which are each multiplied by the exposure variable CAT.

$$\text{logit P}(\mathbf{X}) = \alpha + \beta\text{CAT}$$

$$\underbrace{+\gamma_1\text{AGE}+\gamma_2\text{CHL}+\gamma_3\text{SMK}+\gamma_4\text{ECG}+\gamma_5\text{HPT}}_{\text{main effects}}$$

$$\underbrace{+\ \delta_1\text{CAT}\times\text{CHL}+\ \delta_2\text{CAT}\times\text{HPT}}_{\text{interaction effects}}$$

The model is written as logit P(**X**) equals α plus β times CAT plus the sum of five main effect terms γ_1 times AGE plus γ_2 times CHL and so on up through γ_5 times HPT plus the sum of δ_1 times CAT times CHL plus δ_2 times CAT times HPT. Here the five main effect terms account for the potential confounding effect of the variables AGE through HPT and the two product terms account for the potential interaction effects of CHL and HPT.

Parameters:
α, β, γs, and δs instead of α and βs,

where
β: exposure variable
γs: potential confounders
δs: potential interaction variables

Note that the parameters in this model are denoted as α, β, γs, and δs, whereas previously we denoted all parameters other than the constant α as β_is. We use β, γs, and δs here to distinguish different types of variables in the model. The parameter β indicates the coefficient of the exposure variable, the γs indicate the coefficients of the potential confounders in the model, and the δs indicate the coefficients of the potential interaction variables in the model. This notation for the parameters will be used throughout the remainder of this presentation.

The general *E, V, W* Model

single exposure, controlling for C_1, C_2, . . . , C_p

Analogous to the above example, we now describe the general form of a logistic model, called the *E, V, W* model, that considers the effect of a single exposure controlling for the potential confounding and interaction effects of control variables C_1, C_2, up through C_p.

E, V, W Model

$k = p_1 + p_2 + 1 =$ no. of variables in model
$p_1 =$ no. of potential confounders
$p_2 =$ no. of potential interactions
$1 =$ exposure variable

The general *E, V, W* model contains p_1 plus p_2 plus 1 variables, where p_1 is the number of potential confounders in the model, p_2 is the number of potential interaction terms in the model, and 1 denotes the exposure variable.

CHD EXAMPLE

$p_1 = 5$: AGE, CHL, SMK, ECG, HPT

$p_2 = 2$: CAT × CHL, CAT × HPT

$p_1 + p_2 + 1 = 5 + 2 + 1 = 8$

In the CHD study example above, there are p_1 equals to five potential confounders, namely, the five control variables, and there are p_2 equal to two interaction variables, the first of which is CAT × CHL and the second is CAT × HPT. The total number of variables in the example is, therefore, p_1 plus p_2 plus 1 equals 5 plus 2 plus 1, which equals 8. This corresponds to the model presented earlier, which contained eight variables.

- $V_1, \ldots, V_{p_1}$ are potential confounders
- *V*s are functions of *C*s

In addition to the exposure variable *E*, the general model contains p_1 variables denoted as V_1, V_2 through V_{p_1}. The set of *V*s are functions of the *C*s that are thought to account for confounding in the data. We call the set of these *V*s *potential confounders*.

e.g., $V_1 = C_1, V_2 = (C_2)^2, V_3 = C_1 \times C_3$

For instance, we may have V_1 equal to C_1, V_2 equal to $(C_2)^2$, and V_3 equal to $C_1 \times C_3$.

CHD EXAMPLE

$V_1 =$ AGE, $V_2 =$ CHL, $V_3 =$ SMK, $V_4 =$ ECG, $V_5 =$ HPT

The CHD example above has five *V*s that are the same as the *C*s.

- $W_1, \ldots, W_{p_2}$ are potential effect modifiers
- *W*s are functions of *C*s

Following the *V*s, we define p_2 variables that are product terms of the form *E* times W_1, *E* times W_2, and so on up through *E* times W_{p_2}, where W_1, W_2, through W_{p_2}, denote a set of functions of the *C*s that are *potential effect modifiers* with *E*.

e.g., $W_1 = C_1$, $W_2 = C_1 \times C_3$

For instance, we may have W_1 equal to C_1 and W_2 equal to C_1 times C_3.

CHD EXAMPLE

$W_1 =$ CHL, $W_2 =$ HPT

The CHD example above has two *W*s, namely, CHL and HPT, that go into the model as product terms of the form CAT × CHL and CAT × HPT.

REFERENCES FOR CHOICE OF Vs AND Ws FROM Cs

- Chap. 6: Modeling Strategy Guidelines
- *Epidemiologic Research*, Chap. 21

It is beyond the scope of this chapter to discuss the subtleties involved in the particular choice of the *V*s and *W*s from the *C*s for a given model. More depth is provided in a separate chapter (Chap. 6) on modeling strategies and in Chap. 21 of *Epidemiologic Research* by Kleinbaum, Kupper, and Morgenstern.

Assume: *V*s and *W*s are *C*s or subset of *C*s

EXAMPLE

C_1 = AGE, C_2 = RACE, C_3 = SEX

V_1 = AGE, V_2 = RACE, V_3 = SEX

W_1 = AGE, W_2 = SEX

$p_1 = 3, p_2 = 2, k = p_1 + p_2 + 1 = 6$

In most applications, the *V*s will be the *C*s themselves or some subset of the *C*s and the *W*s will also be the *C*s themselves or some subset thereof. For example, if the *C*s are AGE, RACE, and SEX, then the *V*s may be AGE, RACE, and SEX, and the *W*s may be AGE and SEX, the latter two variables being a subset of the *C*s. Here the number of *V* variables, p_1, equals 3, and the number of *W* variables, p_2, equals 2, so that *k*, which gives the total number of variables in the model, is p_1 plus p_2 plus 1 equals 6.

NOTE
Ws ARE SUBSET OF Vs

EXAMPLE

~~V_1 = AGE, V_2 = RACE~~

~~W_1 = AGE, W_2 = SEX~~

Note, as we describe further in Chap. 6, that you cannot have a *W* in the model that is not also contained in the model as a *V*; that is, *W*s have to be a subset of the *V*s. For instance, we cannot allow a model whose *V*s are AGE and RACE and whose *W*s are AGE and SEX because the SEX variable is not contained in the model as a *V* term.

$$\begin{aligned}\text{logit}\,\text{P}(\mathbf{X}) = \alpha + \beta\text{E} + \gamma_1 V_1 + \gamma_2 V_2 \\ + \cdots + \gamma_{p_1} V_{p_1} + \delta_1 EW_1 \\ + \delta_2 EW_2 + \cdots + \delta_{p_2} EW_{p_2},\end{aligned}$$

A logistic model incorporating this special case containing the *E*, *V*, and *W* variables defined above can be written in logit form as shown here.

where
β = coefficient of *E*
γs = coefficient of *V*s
δs = coefficient of *W*s

Note that β is the coefficient of the single exposure variable *E*, the γs are coefficients of potential confounding variables denoted by the *V*s, and the δs are coefficients of potential interaction effects involving *E* separately with each of the *W*s.

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha + \beta E \\ + \sum_{i=1}^{p_1} \gamma_i V_i + E \sum_{j=1}^{p_2} \delta_j W_j\end{aligned}$$

We can factor out the *E* from each of the interaction terms, so that the model may be more simply written as shown here. This is the form of the model that we will use henceforth in this presentation.

Adjusted odds ratio for $E = 1$ vs. $E = 0$ given $C_1, C_2, \ldots, C_p$ fixed

We now provide for this model an expression for an adjusted odds ratio that describes the effect of the exposure variable on disease status adjusted for the potential confounding and interaction effects of the control variables C_1 through C_p. That is, we give a formula for the risk odds ratio comparing the odds of disease development for exposed vs. unexposed persons, with both groups having the same values for the extraneous factors C_1 through C_p. This formula is derived as a special case of the odds ratio formula for a general logistic model given earlier in our review.

$$\text{ROR} = \exp\left(\beta + \sum_{j=1}^{p_2} \delta_j W_j\right)$$

For our special case, the odds ratio formula takes the form ROR equals e to the quantity β plus the sum from 1 through p_2 of the δ_j times W_j.

- γ_i terms not in formula

Note that β is the coefficient of the exposure variable E, that the δ_j are the coefficients of the interaction terms of the form E times W_j, and that the coefficients γ_i of the main effect variables V_i do not appear in the odds ratio formula.

- Formula assumes E is (0, 1)
- Formula is modified if E has other coding, e.g., (1, −1), (2, 1), ordinal, or interval (see Chap. 3 on coding)

Note also that this formula assumes that the dichotomous variable E is coded as a (0, 1) variable with E equal to 1 for exposed persons and E equal to 0 for unexposed persons. If the coding scheme is different, for example, (1, −1) or (2, 1), or if E is an ordinal or interval variable, then the odds ratio formula needs to be modified. The effect of different coding schemes on the odds ratio formula will be described in Chap. 3.

Interaction:
$$\text{ROR} = \exp\left(\beta + \Sigma\,\boxed{\delta_j W_j}\right)$$

- $\delta_j \neq 0 \Rightarrow$ OR depends on W_j
- Interaction $\Rightarrow$ effect of E differs at different levels of Ws

This odds ratio formula tells us that if our model contains interaction terms, then the odds ratio will involve coefficients of these interaction terms and that, moreover, the value of the odds ratio will be different depending on the values of the W variables involved in the interaction terms as products with E. This property of the OR formula should make sense in that the concept of interaction implies that the effect of one variable, in this case E, is different at different levels of another variable, such as any of the Ws.

- Vs not in OR formula but Vs in model, so OR formula controls confounding:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \Sigma \enclose{circle}{\gamma_i} V_i + E\,\Sigma \enclose{circle}{\delta_j} W_j$$

Although the coefficients of the V terms do not appear in the odds ratio formula, these terms are still part of the fitted model. Thus, the odds ratio formula not only reflects the interaction effects in the model but also controls for the confounding variables in the model.

No interaction:

$$\text{all } \delta_j = 0 \Rightarrow \text{ROR} = \exp\underset{\substack{\uparrow \\ \text{constant}}}{(\beta)}$$

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \underset{\substack{\uparrow \\ \text{confounding} \\ \text{effects adjusted}}}{\sum \gamma_i V_i}$$

In contrast, if the model contains no interaction terms, then, equivalently, all the δ_j coefficients are 0; the odds ratio formula thus reduces to ROR equals to e to β, where β is the coefficient of the exposure variable E. Here, the *odds ratio is a fixed constant*, so that its value does not change with different values of the independent variables. The model in this case reduces to logit P($\mathbf{X}$) equals α plus β times E plus the sum of the main effect terms involving the Vs and contains no product terms. For this model, we can say that e to β represents an odds ratio that *adjusts for the potential confounding effects* of the control variables C_1 through C_p defined in terms of the Vs.

EXAMPLE

The model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta\text{CAT} \underbrace{+\gamma_1\text{AGE} + \gamma_2\text{CHL} + \gamma_3\text{SMK} + \gamma_4\text{ECG} + \gamma_5\text{HPT}}_{\text{main effects}} \underbrace{+ \text{CAT}(\delta_1\text{CHL} + \delta_2\text{HPT})}_{\text{interaction effects}}$$

As an example of the use of the odds ratio formula for the E, V, W model, we return to the CHD study example we described earlier. The CHD study model contained eight independent variables. The model is restated here as logit P($\mathbf{X}$) equals α plus β times CAT plus the sum of five main effect terms plus the sum of two interaction terms.

$$\text{logit P}(X) = \alpha + \beta\text{CAT} \underbrace{+\gamma_1\text{AGE} + \gamma_2\text{CHL} + \gamma_3\text{SMK} + \gamma_4\text{ECG} + \gamma_5\text{HPT}}_{\text{main effects: confounding}} \underbrace{+ \text{CAT}(\delta_1\text{CHL} + \delta_2\text{HPT})}_{\text{product terms: interaction}}$$

The five main effect terms in this model account for the potential confounding effects of the variables AGE through HPT. The two product terms account for the potential interaction effects of CHL and HPT with CAT.

$$\text{ROR} = \exp(\beta + \delta_1\text{CHL} + \delta_2\text{HPT})$$

For this example, the odds ratio formula reduces to the expression ROR equals e to the quantity β plus the sum δ_1 times CHL plus δ_2 times HPT.

EXAMPLE (continued)

$\text{ROR} = \exp(\hat{\beta} + \hat{\delta}_1\text{CHL} + \hat{\delta}_2\text{HPT})$

- varies with values of CHL and HPT

AGE, SMK, and ECG are adjusted for confounding

In using this formula, note that to obtain a numerical value for this odds ratio, not only do we need estimates of the coefficients β and the two δs, but we also need to specify values for the variables CHL and HPT. In other words, once we have fitted the model to obtain estimates of the coefficients, we will get different values for the odds ratio depending on the values that we specify for the interaction variables in our model. Note, also, that although the variables AGE, SMK, and ECG are not contained in the odds ratio expression for this model, the confounding effects of these three variables plus CHL and HPT are being adjusted because the model being fit contains all five control variables as main effect *V* terms.

$n = 609$ white males from Evans County, GA 9-year follow up

Fitted model:

Variable	*Coefficient*
Intercept	$\hat{\alpha} = -4.0497$
CAT	$\hat{\beta} = -12.6894$
AGE	$\hat{\gamma}_1 = 0.0350$
CHL	$\hat{\gamma}_2 = -0.0055$
SMK	$\hat{\gamma}_3 = 0.7732$
ECG	$\hat{\gamma}_4 = 0.3671$
HPT	$\hat{\gamma}_5 = 1.0466$
CAT × CHL	$\hat{\delta}_1 = 0.0692$
CAT × HPT	$\hat{\delta}_2 = -2.3318$

To provide numerical values for the above odds ratio, we will consider a data set of 609 white males from Evans County, Georgia, who were followed for 9 years to determine CHD status. The above model involving CAT, the five *V* variables, and the two *W* variables was fit to this data, and the fitted model is given by the list of coefficients corresponding to the variables listed here.

$$\widehat{\text{ROR}} = \exp(\underbrace{-12.6894}_{\text{exposure coefficient}} + \underbrace{0.0692\text{CHL} - 2.3318\text{ HPT}}_{\text{interaction coefficient}})$$

Based on the above fitted model, the estimated odds ratio for the CAT, CHD association adjusted for the five control variables is given by the expression shown here. Note that this expression involves only the coefficients of the exposure variable CAT and the interaction variables CAT times CHL and CAT times HPT, the latter two coefficients being denoted by δs in the model.

EXAMPLE (continued)

$\widehat{\text{ROR}}$ varies with values of CHL and HPT

effect modifiers

This expression for the odds ratio tells us that we obtain a different value for the estimated odds ratio depending on the values specified for CHL and HPT. As previously mentioned, this should make sense conceptually because CHL and HPT are the only two effect modifiers in the model, and the value of the odds ratio changes as the values of the effect modifiers change.

- CHL = 220, HPT = 1

$$\begin{aligned}\widehat{\text{ROR}} &= \exp[-12.6894 + 0.0692(220)\\ &\quad - 2.3318(1)]\\ &= \exp(0.2028) = \boxed{1.22}\end{aligned}$$

To get a numerical value for the odds ratio, we consider, for example, the specific values CHL equal to 220 and HPT equal to 1. Plugging these into the odds ratio formula, we obtain e to the 0.2028, which equals 1.22.

- CHL = 200, HPT = 0

$$\begin{aligned}\widehat{\text{ROR}} &= \exp[-12.6894 + 0.0692(200)\\ &\quad - 2.3318(0)]\\ &= \exp(1.1506) = \boxed{3.16}\end{aligned}$$

As a second example, we consider CHL equal to 200 and HPT equal to 0. Here, the odds ratio becomes e to 1.1506, which equals 3.16.

CHL = 220, HPT = 1 $\Rightarrow \widehat{\text{ROR}} = 1.22$
CHL = 200, HPT = 0 $\Rightarrow \widehat{\text{ROR}} = 3.16$

controls for the confounding effects of AGE, CHL, SMK, ECG, and HPT

Thus, we see that depending on the values of the effect modifiers we will get different values for the estimated odds ratios. Note that each estimated odds ratio obtained adjusts for the confounding effects of all five control variables because these five variables are contained in the fitted model as *V* variables.

Choice of *W* values depends on investigator

EXAMPLE

TABLE OF POINT ESTIMATES $\widehat{\text{ROR}}$

	HPT = 0	HPT = 1
CHL = 180	0.79	0.08
CHL = 200	3.16	0.31
CHL = 220	12.61	1.22
CHL = 240	50.33	4.89

In general, when faced with an odds ratio expression involving effect modifiers (*W*), the choice of values for the *W* variables depends primarily on the interest of the investigator. Typically, the investigator will choose a range of values for each interaction variable in the odds ratio formula; this choice will lead to a table of estimated odds ratios, such as the one presented here, for a range of CHL values and the two values of HPT. From such a table, together with a table of confidence intervals, the investigator can interpret the exposure–disease relationship.

EXAMPLE

No interaction model for Evans County data (n = 609)

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha &+ \beta\text{CAT}\\ &+ \gamma_1\text{AGE} + \gamma_2\text{CHL}\\ &+ \gamma_3\text{SMK} + \gamma_4\text{ECG}\\ &+ \gamma_5\text{HPT}\end{aligned}$$

As a second example, we consider a model containing no interaction terms from the same Evans County data set of 609 white males. The variables in the model are the exposure variable CAT, and five *V* variables, namely, AGE, CHL, SMK, ECG, and HPT. This model is written in logit form as shown here.

EXAMPLE (continued)

$\widehat{\text{ROR}} = \exp\left(\hat{\beta}\right)$

Because this model contains no interaction terms, the odds ratio expression for the CAT, CHD association is given by e to the $\hat{\beta}$, where $\hat{\beta}$ is the estimated coefficient of the exposure variable CAT.

Fitted model:

Variable	*Coefficient*
Intercept	$\hat{\alpha} = -6.7747$
CAT	$\hat{\beta} = 0.5978$
AGE	$\hat{\gamma}_1 = 0.0322$
CHL	$\hat{\gamma}_2 = 0.0088$
SMK	$\hat{\gamma}_3 = 0.8348$
ECG	$\hat{\gamma}_4 = 0.3695$
HPT	$\hat{\gamma}_5 = 0.4392$

When fitting this no interaction model to the data, we obtain estimates of the model coefficients that are listed here.

$\widehat{\text{ROR}} = \exp(0.5978) = 1.82$

For this fitted model, then, the odds ratio is given by e to the power 0.5978, which equals 1.82. Note that this odds ratio is a fixed number, which should be expected, as there are no interaction terms in the model.

EXAMPLE COMPARISON

	Interaction model	*No interaction model*
Intercept	−4.0497	−6.7747
CAT	−12.6894	0.5978
AGE	0.0350	0.0322
CHL	−0.0055	0.0088
SMK	0.7732	0.8348
ECG	0.3671	0.3695
HPT	1.0466	0.4392
CAT × CHL	0.0692	–
CAT × HPT	−2.3318	–

In comparing the results for the no interaction model just described with those for the model containing interaction terms, we see that the estimated coefficient for any variable contained in both models is different in each model. For instance, the coefficient of CAT in the *no interaction* model is 0.5978, whereas the coefficient of CAT in the *interaction* model is −12.6894. Similarly, the coefficient of AGE in the no interaction model is 0.0322, whereas the coefficient of AGE in the interaction model is 0.0350.

Which model? Requires *strategy*

It should not be surprising to see different values for corresponding coefficients as the two models give a different description of the underlying relationship among the variables. To decide which of these models, or maybe what other model, is more appropriate for this data, we need to use a *strategy* for model selection that includes carrying out tests of significance. A discussion of such a strategy is beyond the scope of this presentation but is described elsewhere (see Chaps. 6 and 7).

This presentation is now complete. We have described important special cases of the logistic model, namely, models for

SUMMARY

1. Introduction

✓ 2. Important Special Cases

- simple analysis
- interaction assessment involving two variables
- assessment of potential confounding and interaction effects of several covariates

We suggest that you review the material covered here by reading the detailed outline that follows. Then do the practice exercises and test.

3. Computing the Odds Ratio

All of the special cases in this presentation involved a (0, 1) exposure variable. In the next chapter, we consider how the odds ratio formula is modified for other codings of single exposures and also examine several exposure variables in the same model, controlling for potential confounders and effect modifiers.

Detailed Outline

I. Overview (page 45)

A. Focus:

- Simple analysis
- Multiplicative interaction
- Controlling several confounders and effect modifiers

B. Logistic model formula when $\mathbf{X} = (X_1, X_2, \ldots, X_k)$:

$$\mathrm{P}(\mathbf{X}) = \frac{1}{1 + \mathrm{e}^{-\left(\alpha + \sum\limits_{i=1}^{k} \beta_i X_i\right)}}.$$

C. Logit form of logistic model:

$$\text{logit } \mathrm{P}(\mathbf{X}) = \alpha + \sum_{i=1}^{k} \beta_i X_i.$$

D. General odds ratio formula:

$$\mathrm{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} = \mathrm{e}^{\sum\limits_{i=1}^{k} \beta_i (X_{1i} - X_{0i})} = \prod_{i=1}^{k} \mathrm{e}^{\beta_i (X_{1i} - X_{0i})}.$$

II. Special case – Simple analysis (pages 46–49)

A. The model:

$$\mathrm{P}(\mathbf{X}) = \frac{1}{1 + \mathrm{e}^{-(\alpha + \beta_1 E)}}$$

B. Logit form of the model:

$$\text{logit } \mathrm{P}(\mathbf{X}) = \alpha + \beta_1 E$$

C. Odds ratio for the model: $\mathrm{ROR} = \exp(\beta_1)$

D. Null hypothesis of no *E, D* effect: H_0: $\beta_1 = 0$.

E. The estimated odds ratio $\exp(\hat{\beta})$ is computationally equal to *ad/bc* where *a, b, c,* and *d* are the cell frequencies within the four-fold table for simple analysis.

III. Assessing multiplicative interaction (pages 49–55)

A. Definition of no interaction on a multiplicative scale: $\mathrm{OR}_{11} = \mathrm{OR}_{10} \times \mathrm{OR}_{01}$, where OR_{AB} denotes the odds ratio that compares a person in category *A* of one factor and category *B* of a second factor with a person in referent categories 0 of both factors, where *A* takes on the values 0 or 1 and *B* takes on the values 0 or 1.

B. Conceptual interpretation of no interaction formula: The effect of both variables *A* and *B* acting together is the same as the combined effect of each variable acting separately.

C. Examples of no interaction and interaction on a multiplicative scale.

D. A logistic model that allows for the assessment of multiplicative interaction:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1 A + \beta_2 B + \beta_3 A \times B$$

E. The relationship of β_3 to the odds ratios in the no interaction formula above:

$$\beta_3 = \ln\left(\frac{\text{OR}_{11}}{\text{OR}_{10} \times \text{OR}_{01}}\right)$$

F. The null hypothesis of no interaction in the above two factor model: H_0: $\beta_3 = 0$.

IV. The *E, V, W* model – A general model containing a (0, 1) exposure and potential confounders and effect modifiers (pages 55–64)

A. Specification of variables in the model: start with $E, C_1, C_2, \ldots, C_p$; then specify potential confounders $V_1, V_2, \ldots, V_{p_1}$, which are functions of the *C*s, and potential interaction variables (i.e., effect modifiers) $W_1, W_2, \ldots, W_{p_2}$, which are also functions of the *C*s and go into the model as product terms with *E*, i.e., $E \times W_j$.

B. The *E, V, W* model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \sum_{i=1}^{p_1} \gamma_i V_i + E \sum_{j=1}^{p_2} \delta_j W_j$$

C. Odds ratio formula for the *E, V, W* model, where *E* is a (0, 1) variable:

$$\text{ROR}_{E\,=\,1\text{ vs. }E\,=\,0} = \exp\left(\beta + \sum_{j=1}^{p_2} \delta_j W_j\right)$$

D. Odds ratio formula for *E, V, W* model if no interaction: ROR $= \exp(\beta)$.

E. Examples of the *E, V, W* model: with interaction and without interaction

Practice Exercises

True or False (Circle T or F)

T F 1. A logistic model for a simple analysis involving a (0, 1) exposure variable is given by logit P(**X**) = $\alpha + \beta E$, where E denotes the (0, 1) exposure variable.

T F 2. The odds ratio for the exposure–disease relationship in a logistic model for a simple analysis involving a (0, 1) exposure variable is given by β, where β is the coefficient of the exposure variable.

T F 3. The null hypothesis of no exposure–disease effect in a logistic model for a simple analysis is given by H_0: $\beta = 1$, where β is the coefficient of the exposure variable.

T F 4. The log of the estimated coefficient of a (0, 1) exposure variable in a logistic model for simple analysis is equal to ad/bc, where a, b, c, and d are the cell frequencies in the corresponding fourfold table for simple analysis.

T F 5. Given the model logit P(**X**) = $\alpha + \beta E$, where E denotes a (0, 1) exposure variable, the *risk* for exposed persons ($E = 1$) is expressible as e^{β}.

T F 6. Given the model logit P(**X**) = $\alpha + \beta E$, as in Exercise 5, the *odds* of getting the disease for exposed persons ($E = 1$) is given by $e^{\alpha+\beta}$.

T F 7. A logistic model that incorporates a multiplicative interaction effect involving two (0, 1) independent variables X_1 and X_2 is given by logit P(**X**) = $\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$.

T F 8. An equation that describes "no interaction on a multiplicative scale" is given by $\mathrm{OR}_{11} = \mathrm{OR}_{10}/\mathrm{OR}_{01}$.

T F 9. Given the model logit P(**X**) = $\alpha + \beta E + \gamma$SMK + $\delta E \times$ SMK, where E is a (0, 1) exposure variable and SMK is a (0, 1) variable for smoking status, the null hypothesis for a test of no interaction on a multiplicative scale is given by H_0: $\delta = 0$.

T F 10. For the model in Exercise 9, the odds ratio that describes the exposure disease effect controlling for smoking is given by $\exp(\beta + \delta)$.

T F 11. Given an exposure variable E and control variables AGE, SBP, and CHL, suppose it is of interest to fit a model that adjusts for the potential confounding effects of all three control variables considered as main effect terms and for the potential interaction effects with E of all three control variables. Then the logit form of a model that describes this situation is given by logit $\mathrm{P}(\mathbf{X}) = \alpha + \beta E + \gamma_1\mathrm{AGE} + \gamma_2\mathrm{SBP} + \gamma_3\mathrm{CHL} + \delta_1\mathrm{AGE} \times \mathrm{SBP} + \delta_2\mathrm{AGE} \times \mathrm{CHL} + \delta_3\mathrm{SBP} \times \mathrm{CHL}$.

T F 12. Given a logistic model of the form logit $\mathrm{P}(\mathbf{X}) = \alpha + \beta E + \gamma_1\mathrm{AGE} + \gamma_2\mathrm{SBP} + \gamma_3\mathrm{CHL}$, where E is a (0, 1) exposure variable, the odds ratio for the effect of E adjusted for the confounding of AGE, CHL, and SBP is given by $\exp(\beta)$.

T F 13. If a logistic model contains interaction terms expressible as products of the form EW_j where W_j are potential effect modifiers, then the value of the odds ratio for the E, D relationship will be different, depending on the values specified for the W_j variables.

T F 14. Given the model logit $\mathrm{P}(\mathbf{X}) = \alpha + \beta E + \gamma_1\mathrm{SMK} + \gamma_2\mathrm{SBP}$, where E and SMK are (0, 1) variables, and SBP is continuous, then the odds ratio for estimating the effect of SMK on the disease, controlling for E and SBP is given by $\exp(\gamma_1)$.

T F 15. Given E, C_1, and C_2, and letting $V_1 = C_1 = W_1$ and $V_2 = C_2 = W_2$, then the corresponding logistic model is given by logit $\mathrm{P}(\mathbf{X}) = \alpha + \beta E + \gamma_1 C_1 + \gamma_2 C_2 + E(\delta_1 C_1 + \delta_2 C_2)$.

T F 16. For the model in Exercise 15, if $C_1 = 20$ and $C_2 = 5$, then the odds ratio for the E, D relationship has the form $\exp(\beta + 20\delta_1 + 5\delta_2)$.

Test

True or False (Circle T or F)

T F 1. Given the simple analysis model, logit $P(\mathbf{X}) = \phi + \psi Q$, where ϕ and ψ are unknown parameters and Q is a (0, 1) exposure variable, the odds ratio for describing the exposure–disease relationship is given by $\exp(\phi)$.

T F 2. Given the model logit $P(\mathbf{X}) = \alpha + \beta E$, where E denotes a (0, 1) exposure variable, the *risk* for unexposed persons ($E = 0$) is expressible as $1/\exp(-\alpha)$.

T F 3. Given the model in Question 2, the *odds* of getting the disease for unexposed persons ($E = 0$) is given by $\exp(\alpha)$.

T F 4. Given the model logit $P(\mathbf{X}) = \phi + \psi\text{HPT} + \rho\text{ECG} + \pi\text{HPT} \times \text{ECG}$, where HPT is a (0, 1) exposure variable denoting hypertension status and ECG is a (0, 1) variable for electrocardiogram status, the null hypothesis for a test of no interaction on a multiplicative scale is given by H_0: $\exp(\pi) = 1$.

T F 5. For the model in Question 4, the odds ratio that describes the effect of HPT on disease status, controlling for ECG, is given by $\exp(\psi + \pi\text{ECG})$.

T F 6. Given the model logit $P(\mathbf{X}) = \alpha + \beta E + \phi\text{HPT} + \psi\text{ECG}$, where E, HPT, and ECG are (0, 1) variables, then the odds ratio for estimating the effect of ECG on the disease, controlling for E and HPT, is given by $\exp(\psi)$.

T F 7. Given E, C_1, and C_2, and letting $V_1 = C_1 = W_1$, $V_2 = (C_1)^2$, and $V_3 = C_2$, then the corresponding logistic model is given by logit $P(\mathbf{X}) = \alpha + \beta E + \gamma_1 C_1 + \gamma_2 {C_1}^2 + \gamma_3 C_2 + \delta E\text{C}_1$.

T F 8. For the model in Question 7, if $\text{C}_1 = 5$ and $C_2 = 20$, then the odds ratio for the E, D relationship has the form $\exp(\beta + 20\delta)$.

Consider a 1-year follow-up study of bisexual males to assess the relationship of behavioral risk factors to the acquisition of HIV infection. Study subjects were all in the 20–30 age range and were enrolled if they tested HIV negative and had claimed not to have engaged in "high-risk" sexual activity for at least 3 months. The outcome variable is HIV status at 1 year, a (0, 1) variable, where a subject gets the value 1 if HIV positive and 0 if HIV negative at 1 year after start of follow-up. Four risk factors were considered: consistent and correct condom use (CON), a (0, 1) variable; having one or more sex partners in high-risk groups (PAR), also a (0, 1) variable; the number of sexual partners (NP); and the average number of sexual contacts per month (ASCM). The primary purpose of this study was to determine the effectiveness of consistent and correct condom use in preventing the acquisition of HIV infection, controlling for the other variables. Thus, the variable CON is considered the exposure variable, and the variables PAR, NP, and ASCM are potential confounders and potential effect modifiers.

9. Within the above study framework, state the logit form of a logistic model for assessing the effect of CON on HIV acquisition, controlling for each of the other three risk factors as both potential confounders and potential effect modifiers. (Note: In defining your model, *only* use interaction terms that are two-way products of the form $E \times W$, where E is the exposure variable and W is an effect modifier.)
10. Using the model in Question 9, give an expression for the odds ratio that compares an exposed person (CON = 1) with an unexposed person (CON = 0) who has the same values for PAR, NP, and ASCM.

Answers to Practice Exercises

1. T
2. F: OR $= e^{\beta}$
3. F: H_0: $\beta = 0$
4. F: $e^{\beta} = ad/bc$
5. F: risk for $E = 1$ is $1/[1 + e^{-(\alpha+\beta)}]$
6. T
7. T
8. F: $OR_{11} = OR_{10} \times OR_{01}$
9. T
10. F: OR $= \exp(\beta + \delta\text{SMK})$
11. F: interaction terms should be $E \times$ AGE, $E \times$ SBP, and $E \times$ CHL
12. T
13. T
14. T
15. T
16. T

3 Computing the Odds Ratio in Logistic Regression

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_3, © Springer Science+Business Media, LLC 2010

Introduction

In this chapter, the ***E, V, W model*** is extended to consider other coding schemes for a single exposure variable, including ordinal and interval exposures. The model is further extended to allow for several exposure variables. The formula for the odds ratio is provided for each extension, and examples are used to illustrate the formula.

Abbreviated Outline

The outline below gives the user a preview of the material covered by the presentation. Together with the objectives, this outline offers the user an overview of the content of this module. A detailed outline for review purposes follows the presentation.

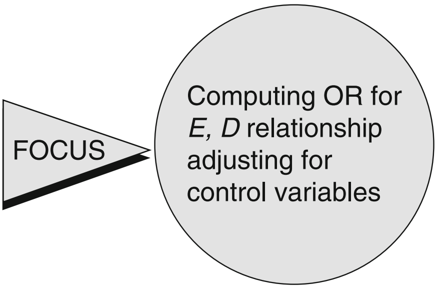
Objectives

Upon completing this chapter, the learner should be able to:

1. Given a logistic model for a study situation involving a single exposure variable and several control variables, compute or recognize the expression for the odds ratio for the effect of exposure on disease status that adjusts for the confounding and interaction effects of functions of control variables:
    a. When the exposure variable is dichotomous and coded (*a*, *b*) for any two numbers *a* and *b*
    b. When the exposure variable is ordinal and two exposure values are specified
    c. When the exposure variable is continuous and two exposure values are specified
2. Given a study situation involving a single nominal exposure variable with more than two (i.e., polytomous) categories, state or recognize a logistic model that allows for the assessment of the exposure–disease relationship controlling for potential confounding and assuming no interaction.
3. Given a study situation involving a single nominal exposure variable with more than two categories, compute or recognize the expression for the odds ratio that compares two categories of exposure status, controlling for the confounding effects of control variables and assuming no interaction.
4. Given a study situation involving several distinct exposure variables, state or recognize a logistic model that allows for the assessment of the joint effects of the exposure variables on disease controlling for the confounding effects of control variables and assuming no interaction.
5. Given a study situation involving several distinct exposure variables, state or recognize a logistic model that allows for the assessment of the joint effects of the exposure variables on disease controlling for the confounding and interaction effects of control variables.

Presentation

I. Overview


FOCUS
Computing OR for *E, D* relationship adjusting for control variables

This presentation describes how to compute the odds ratio for special cases of the general logistic model involving one or more exposure variables. We focus on models that allow for the assessment of an exposure–disease relationship that adjusts for the potential confounding and/or effect modifying effects of control variables.

- Dichotomous *E* – arbitrary coding
- Ordinal or interval *E*
- Polytomous *E*
- Several *E*s

In particular, we consider dichotomous exposure variables with arbitrary coding, that is, the coding of exposure may be other than (0, 1). We also consider single exposures that are ordinal or interval scaled variables. And, finally, we consider models involving several exposures, a special case of which involves a single polytomous exposure.

Chapter 2 – *E, V, W* model:

- (0, 1) exposure
- Confounders
- Effect modifiers

In the previous chapter we described the logit form and odds ratio expression for the *E, V, W* logistic model, where we considered a single (0, 1) exposure variable and we allowed the model to control several potential confounders and effect modifiers.

The variables in the *E, V, W* model:

E:	(0, 1) exposure
*C*s:	control variables
*V*s:	potential confounders
*W*s:	potential effect modifiers (i.e., go into model as $E \times W$)

Recall that in defining the *E, V, W* model, we start with a single dichotomous (0, 1) exposure variable, *E*, and *p* control variables C_1, C_2, and so on, up through C_p. We then define a set of potential confounder variables, which are denoted as *V*s. These *V*s are functions of the *C*s that are thought to account for confounding in the data. We then define a set of potential effect modifiers, which are denoted as *W*s. Each of the *W*s goes into the model as product term with *E*.

The *E, V, W* model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \sum_{i=1}^{p_1} \gamma_i V_i + E \sum_{j=1}^{p_2} \delta_j W_j$$

The *logit form* of the *E, V, W* model is shown here. Note that β is the coefficient of the single exposure variable *E*, the gammas (γs) are coefficients of potential confounding variables denoted by the *V*s, and the deltas (δs) are coefficients of potential interaction effects involving *E* separately with each of the *W*s.

Adjusted odds ratio for effect of E adjusted for Cs:

$$\text{ROR}_{E=1 \text{ vs. } E=0} = \exp\left(\beta + \sum_{j=1}^{p_2} \delta_j W_j\right)$$

(γ_i terms not in formula)

For this model, the formula for the *adjusted odds ratio* for the effect of the exposure variable on disease status adjusted for the potential confounding and interaction effects of the Cs is shown here. This formula takes the form e to the quantity β plus the sum of terms of the form δ_j times W_j. Note that the coefficients γ_i of the main effect variables V_i do not appear in the odds ratio formula.

II. Odds Ratio for Other Codings of a Dichotomous *E*

Need to modify OR formula if coding of E is not (0, 1)

Note that this odds ratio formula assumes that the dichotomous variable E is coded as a (0, 1) *variable with E equal to* 1 *when exposed and E equal to* 0 *when unexposed*. If the coding scheme is different – for example, (−1, 1) or (2, 1), or if E is an ordinal or interval variable – then the odds ratio formula needs to be modified.

Focus: ✓ dichotomous
ordinal
interval

We now consider other coding schemes for dichotomous variables. Later, we also consider coding schemes for ordinal and interval variables.

$$E = \begin{cases} a & \text{if exposed} \\ b & \text{if unexposed} \end{cases}$$

$$\text{ROR}_{E=a \text{ vs. } E=b} = \exp\left[(a-b)\beta + (a-b)\sum_{j=1}^{p_2} \delta_j W_j\right]$$

Suppose E is coded to take on the value a if exposed and b if unexposed. Then, it follows from the general odds ratio formula that ROR equals e to the quantity $(a - b)$ times β plus $(a - b)$ times the sum of the δ_j times the W_j.

EXAMPLES

(A) $a = 1, b = 0 \Rightarrow \underbrace{(\text{a}-\text{b})} = (1-0) = 1$

$$\text{ROR} = \exp(1\beta + 1\sum \delta_j W_j)$$

(B) $a = 1, b = -1 \Rightarrow \underbrace{(\text{a}-\text{b})} = (1-[-1]) = 2$

$$\text{ROR} = \exp(2\beta + 2\sum \delta_j W_j)$$

(C) $a = 100, b = 0 \Rightarrow \underbrace{(\text{a}-\text{b})} = (100-0) = 100$

$$\text{ROR} = \exp(100\beta + 100\sum \delta_j W_j)$$

For example, if a equals 1 and b equals 0, then we are using the (0, 1) coding scheme described earlier. It follows that a minus b equals 1 minus 0, or 1, so that the ROR expression is e to the β plus the sum of the δ_j times the W_j. We have previously given this expression for (0, 1) coding.

In contrast, if a equals 1 and b equals −1, then a minus b equals 1 minus −1, which is 2, so the odds ratio expression changes to e to the quantity 2 times β plus 2 times the sum of the δ_j times the W_j.

As a third example, suppose a equals 100 and b equals 0, then a minus b equals 100, so the odds ratio expression changes to e to the quantity 100 times β plus 100 times the sum of the δ_j times the W_j.

Coding	$\widehat{\text{ROR}}$
(A) $a=1, b=0$	$\widehat{\text{ROR}}_\text{A} = \exp\left(\hat{\beta}_\text{A} + \sum_{j=1}^{p_2} \hat{\delta}_{j\text{A}} W_j\right)$
(B) $a=1, b=-1$	$\widehat{\text{ROR}}_\text{B} = \exp\left(2\hat{\beta}_\text{B} + \sum_{j=1}^{p_2} 2\hat{\delta}_{j\text{B}} W_j\right)$
(C) $a=100, b=0$	$\widehat{\text{ROR}}_\text{C} = \exp\left(100\hat{\beta}_\text{C} + \sum_{j=1}^{p_2} 100\,\hat{\delta}_{j\text{C}} W_j\right)$

same value although different codings

different values for different codings

$\widehat{\text{ROR}}_\text{A} = \widehat{\text{ROR}}_\text{B} = \widehat{\text{ROR}}_\text{C}$

$\hat{\beta}_\text{A} \neq \hat{\beta}_\text{B} \neq \hat{\beta}_\text{C}$

$\hat{\delta}_{jA} \neq \hat{\delta}_{jB} \neq \hat{\delta}_{jC}$

Thus, depending on the coding scheme for E, the odds ratio will be calculated differently. Nevertheless, even though $\hat{\beta}$ and the $\hat{\delta}_j$ will be different for different coding schemes, the final odds ratio value will be the same as long as the correct formula is used for the corresponding coding scheme.

As shown here for the three examples above, which are labeled A, B, and C, the three computed odds ratios will be the same, even though the estimates $\hat{\beta}$ and $\hat{\delta}_j$ used to compute these odds ratios will be different for different codings.

EXAMPLE: No Interaction Model

Evans County follow-up study:
$n = 609$ white males
$D =$ CHD status
$E =$ CAT, dichotomous
$V_1 =$ AGE, $V_2 =$ CHL, $V_3 =$ SMK,
$V_4 =$ ECG, $V_5 =$ HPT

As a numerical example, we consider a model that contains no interaction terms from a data set of 609 white males from Evans County, Georgia. The study is a follow-up study to determine the development of coronary heart disease (CHD) over 9 years of follow-up. The variables in the model are CAT, a dichotomous exposure variable, and five V variables, namely, AGE, CHL, SMK, ECG, and HPT.

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha &+ \beta\text{CAT} + \gamma_1\text{AGE}\\ &+ \gamma_2\text{CHL} + \gamma_3\text{SMK}\\ &+ \gamma_4\text{ECG} + \gamma_5\text{HPT}\end{aligned}$$

This model is written in *logit form* as logit P($\mathbf{X}$) equals α plus β times CAT plus the sum of five main effect terms γ_1 times AGE plus γ_2 times CHL, and so on up through γ_5 times HPT.

CAT: $\underbrace{(0,\ 1)}_{\downarrow}$ vs. other codings

We first describe the results from fitting this model when CAT is coded as a (0, 1) variable. Then, we contrast these results with other codings of CAT.

$$\widehat{\text{ROR}} = \exp\left(\hat{\beta}\right)$$

Because this model contains no interaction terms and CAT is coded as (0, 1), the odds ratio expression for the CAT, CHD association is given by e to $\hat{\beta}$, where $\hat{\beta}$ is the estimated coefficient of the exposure variable CAT.

EXAMPLE (continued)

(0, 1) coding for CAT

Variable	*Coefficient*
Intercept	$\hat{\alpha} = -6.7747$
CAT	$\hat{\beta} = 0.5978$
AGE	$\hat{\gamma}_1 = 0.0322$
CHL	$\hat{\gamma}_2 = 0.0088$
SMK	$\hat{\gamma}_3 = 0.8348$
ECG	$\hat{\gamma}_4 = 0.3695$
HPT	$\hat{\gamma}_5 = 0.4392$

Fitting this no interaction model to the data, we obtain the estimates listed here.

$\widehat{\text{ROR}} = \exp(0.5978) = \boxed{1.82}$
No interaction model: ROR fixed

For this fitted model, then, the odds ratio is given by e to the power 0.5978, which equals 1.82. Notice that, as should be expected, this odds ratio is a fixed number as there are no interaction terms in the model.

$(-1, 1)$ coding for CAT:

$$\hat{\beta} = 0.2989 = \left(\frac{0.5978}{2}\right)$$

$$\begin{aligned}\widehat{\text{ROR}} = \exp\left(2\hat{\beta}\right) &= \exp(2 \times 0.2989)\\ &= \exp(0.5978)\\ &= 1.82\end{aligned}$$

same $\widehat{\text{ROR}}$ as for (0, 1) coding

Note. $\widehat{\text{ROR}} \neq \exp(0.2989) = \underset{\substack{\uparrow\\ \text{incorrect value}}}{1.35}$

Now, if we consider the same data set and the same model, except that the coding of CAT is (−1, 1) instead of (0, 1), the coefficient $\hat{\beta}$ of CAT becomes 0.2989, which is one-half of 0.5978. Thus, for this coding scheme, the odds ratio is computed as e to 2 times the corresponding $\hat{\beta}$ of 0.2989, which is the same as e to 0.5978, or 1.82. We see that, regardless of the coding scheme used, the final odds ratio result is the same, as long as the correct odds ratio formula is used. In contrast, it would be incorrect to use the (−1, 1) coding scheme and then compute the odds ratio as e to 0.2989.

III. Odds Ratio for Arbitrary Coding of *E*

Model:

$$\overset{\substack{\text{dichotomous, ordinal or interval}\\ \downarrow}}{}$$

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \sum_{i=1}^{p_1} \gamma_i V_i + E \sum_{j=1}^{p_2} \delta_j W_j$$

We now consider the odds ratio formula for any single exposure variable *E*, whether *dichotomous, ordinal*, or *interval*, controlling for a collection of *C* variables in the context of an *E, V, W* model shown again here. That is, we allow the variable *E* to be defined arbitrarily of interest.

E^*(group 1) vs. E^{**} (group 2)

To obtain an odds ratio for such a generally defined E, we need to specify two values of E to be compared. We denote the two values of interest as E^* and E^{**}. We need to specify two values because an odds ratio requires the *comparison of two groups* – in this case two levels of the exposure variable E – even when the exposure variable can take on more than two values, as when E is ordinal or interval.

$$\text{ROR}_{E^* \text{ vs. } E^{**}} = \exp\left[(E^* - E^{**})\beta + (E^* - E^{**})\sum_{j=1}^{p_2} \delta_j W_j\right]$$

Same as

$$\text{ROR}_{E=a \text{ vs. } E=b} = \exp\left[(a-b)\beta + (a-b)\sum_{j=1}^{p_2} \delta_j W_j\right]$$

The odds ratio formula for E^* vs. E^{**}, equals e to the quantity (E^* – E^{**}) times β plus (E^* – E^{**}) times the sum of the δ_j times W_j. This is essentially the same formula as previously given for dichotomous E, except that here, several different odds ratios can be computed as the choice of E^* and E^{**} ranges over the possible values of E.

EXAMPLE

$E = \text{SSU} = \text{social support status (0–5)}$

We illustrate this formula with several examples. First, suppose E gives social support status as denoted by SSU, which is an index ranging from 0 to 5, where 0 denotes a person without any social support and 5 denotes a person with the maximum social support possible.

(A) $\text{SSU}^* = 5$ vs. $\text{SSU}^{**} = 0$

$$\begin{aligned}\text{ROR}_{5,0} &= \exp\left[(\text{SSU}^* - \text{SSU}^{**})\beta + (\text{SSU}^* - \text{SSU}^{**})\Sigma\delta_j W_j\right]\\ &= \exp\left[(5-0)\beta + (5-0)\Sigma\delta_j W_j\right]\\ &= \exp(5\beta + 5\Sigma\delta_j W_j)\end{aligned}$$

To obtain an odds ratio involving *social support status (SSU)*, in the context of our *E, V, W* model, we need to specify two values of E. One such pair of values is SSU^* equals 5 and SSU^{**} equals 0, which compares the odds for persons who have the highest amount of social support with the odds for persons who have the lowest amount of social support. For this choice, the odds ratio expression becomes e to the quantity (5 – 0) times β plus (5 – 0) times the sum of the δ_j times W_j, which simplifies to e to 5β plus 5 times the sum of the δ_j times W_j.

(B) $\text{SSU}^* = 3$ vs. $\text{SSU}^{**} = 1$

$$\begin{aligned}\text{ROR}_{3,1} &= \exp\left[(3-1)\beta + (3-1)\textstyle\sum\delta_j W_j\right]\\ &= \exp(2\beta + 2\textstyle\sum\delta_j W_j)\end{aligned}$$

Similarly, if SSU^* equals 3 and SSU^{**} equals 1, then the odds ratio becomes e to the quantity (3 – 1) times β plus (3 – 1) times the sum of the δ_j times W_j, which simplifies to e to 2β plus 2 times the sum of the δ_j times W_j.

EXAMPLE (continued)

(C) $\text{SSU}^* = 4$ vs. $\text{SSU}^{**} = 2$

$$\begin{aligned}\text{ROR}_{4,2} &= \exp\left[(4-2)\beta + (4-2)\textstyle\sum \delta_j W_j\right]\\ &= \exp\left(2\beta + 2\textstyle\sum \delta_j W_j\right)\end{aligned}$$

Note. ROR depends on the difference $(E^* - E^{**})$, e.g., $(3 - 1) = (4 - 2) = 2$

Note that if SSU^* equals 4 and SSU^{**} equals 2, then the odds ratio expression becomes 2β plus 2 times the sum of the δ_j times W_j, which is the same expression as obtained when SSU^* equals 3 and SSU^{**} equals 1. This occurs because the odds ratio depends on the difference between E^* and E^{**}, which in this case is 2, regardless of the specific values of E^* and E^{**}.

EXAMPLE

$E = \text{SBP}$ = systolic blood pressure (interval)

(A) $\text{SBP}^* = 160$ vs. $\text{SBP}^{**} = 120$

$$\begin{aligned}\text{ROR}_{160,120} &= \exp\left[(\text{SBP}^* - \text{SBP}^{**})\beta\right.\\ &\qquad \left. + (\text{SBP}^* - \text{SBP}^{**})\textstyle\sum \delta_j W_j\right]\\ &= \exp\left[(160-120)\beta\right.\\ &\qquad \left. + (160-120)\textstyle\sum \delta_j W_j\right]\\ &= \exp\left(40\beta + 40\textstyle\sum \delta_j W_j\right)\end{aligned}$$

(B) $\text{SBP}^* = 200$ vs. $\text{SBP}^{**} = 120$

$$\begin{aligned}\text{ROR}_{200,120} &= \exp\left[(200-120)\beta + (200-120)\textstyle\sum \delta_j W_j\right]\\ &= \exp\left(80\beta + 80\textstyle\sum \delta_j W_j\right)\end{aligned}$$

As another illustration, suppose E is the interval variable systolic blood pressure denoted by SBP. Again, to obtain an odds ratio, we must specify two values of E to compare. For instance, if SBP^* equals 160 and SBP^{**} equals 120, then the odds ratio expression becomes ROR equals e to the quantity (160 − 120) times β plus (160 − 120) times the sum of the δ_j times W_j, which simplifies to 40 times β plus 40 times the sum of the δ_j times W_j.

Or if SBP^* equals 200 and SBP^{**} equals 120, then the odds ratio expression becomes ROR equals e to the 80 times β plus 80 times the sum of the γ_j times W_j.

No interaction:

$\text{ROR}_{E^* \text{ vs. } E^{**}} = \exp\left[(\text{E}^* - \text{E}^{**})\beta\right]$

If $(E^* - E^{**}) = 1$, then ROR $= \exp(\beta)$

e.g., $E^* = 1$ vs. $E^{**} = 0$

or $E^* = 2$ vs. $E^{**} = 1$

Note that in the no interaction case, the odds ratio formula for a general exposure variable E reduces to e to the quantity $(E^* - E^{**})$ times β. This is not equal to e to the β unless the difference $(E^* - E^{**})$ equals 1, as, for example, if E^* equals 1 and E^{**} equals 0, or E^* equals 2 and E^{**} equals 1.

EXAMPLE

$E = \text{SBP}$

$\text{ROR} = \exp(\beta) \Rightarrow (\text{SBP}^* - \text{SBP}^{**}) = 1$

not interesting ↑

Choice of SBP:

Clinically meaningful categories, e.g., $\text{SBP}^* = 160$, $\text{SBP}^* = 120$

Strategy: Use quintiles of SBP

Quintile #	1	2	3	4	5
Mean or median	120	140	160	180	200

Thus, if E denotes SBP, then the quantity e to β gives the odds ratio for comparing any two groups that differ by one unit of SBP. A one unit difference in SBP is not typically of interest, however. Rather, a typical choice of SBP values to be compared represent clinically meaningful categories of blood pressure, as previously illustrated, for example, by SBP^* equals 160 and SBP^{**} equals 120.

One possible strategy for choosing values of SBP^* and SBP^{**} is to categorize the distribution of SBP values in our data into clinically meaningful categories, say, quintiles. Then, using the mean or median SBP in each quintile, we can compute odds ratios comparing all possible pairs of mean or median SBP values.

EXAMPLE (continued)

SBP^*	SBP^{**}	OR
200	120	✓
200	140	✓
200	160	✓
200	180	✓
180	120	✓
180	140	✓
180	160	✓
160	140	✓
160	120	✓
140	120	✓

For instance, suppose the medians of each quintile are 120, 140, 160, 180, and 200. Then odds ratios can be computed comparing SBP^* equal to 200 with SBP^{**} equal to 120, followed by comparing SBP^* equal to 200 with SBP^{**} equal to 140, and so on until all possible pairs of odds ratios are computed. We would then have a table of odds ratios to consider for assessing the relationship of SBP to the disease outcome variable. The check marks in the table shown here indicate pairs of odds ratios that compare values of SBP^* and SBP^{**}.

IV. The Model and Odds Ratio for a Nominal Exposure Variable (No Interaction Case)

Several exposures: $E_1, E_2, \ldots, E_q$

- Model
- Odds ratio

The final special case of the logistic model that we will consider expands the E, V, W model to allow for several exposure variables. That is, instead of having a single E in the model, we will allow several Es, which we denote by E_1, E_2, and so on up through E_q. In describing such a model, we consider some examples and then give a general model formula and a general expression for the odds ratio.

Nominal variable: $>$ 2 categories

e.g., ✓ occupational status in four groups

~~SSU (0 – 5) ordinal~~

First, suppose we have a single nominal exposure variable of interest; that is, instead of being dichotomous, the exposure contains more than two categories that are not orderable. An example is a variable such as occupational status, which is denoted in general as OCC, but divided into four groupings or occupational types. In contrast, a variable like social support, which we previously denoted as SSU and takes on discrete values ordered from 0 to 5, is an ordinal variable.

k categories $\Rightarrow$ $k - 1$ dummy variables
$E_1, E_2, \ldots, E_{k-1}$

When considering nominal variables in a logistic model, we use dummy variables to distinguish the different categories of the variable. If the model contains an intercept term α, then we use $k - 1$ dummy variables E_1, E_2, and so on up to E_{k-1} to distinguish among k categories.

EXAMPLE

$E = \text{OCC with } k = 4 \Rightarrow k - 1 = 3$
$\text{OCC}_1, \text{OCC}_2, \text{OCC}_3$

$$\text{where OCC}_i = \begin{cases} 1 & \text{if category } i \\ 0 & \text{if otherwise} \end{cases}$$

for $i = 1, 2, 3$ (referent: category 4)

So, for example, with occupational status, we define three dummy variables OCC_1, OCC_2, and OCC_3 to reflect four occupational categories, where OCC_i is defined to take on the value 1 for a person in the ith occupational category and 0 otherwise, for i ranging from 1 to 3. Note that for this choice of dummy variables, the referent group is the fourth occupational category, for which $\text{OCC}_1 = \text{OCC}_2 = \text{OCC}_3 = 0$.

No interaction model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1 E_1 + \beta_2 E_2 + \ldots + \beta_{k-1} E_{k-1} + \sum_{i=1}^{p_1} \gamma_i V_i$$

A no interaction model for a nominal exposure variable with k categories then takes the form logit P($\mathbf{X}$) equals α plus β_1 times E_1 plus β_2 times E_2 and so on up to β_{k-1} times E_{k-1} plus the usual set of V terms, where the E_i are the dummy variables described above.

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1 \text{OCC}_1 + \beta_2 \text{OCC}_2 + \beta_3 \text{OCC}_3 + \sum_{i=1}^{p_1} \gamma_i V_i$$

The corresponding model for four occupational status categories then becomes logit P($\mathbf{X}$) equals α plus β_1 times OCC_1 plus β_2 times OCC_2 plus β_3 times OCC_3 plus the V terms.

Specify $\mathbf{E}^*$ and $\mathbf{E}^{**}$ in terms of $k - 1$ dummy variables where

$$\mathbf{E} = (E_1, E_2, \ldots, E_{k-1})$$

To obtain an odds ratio from the above model, we need to specify two categories $\mathbf{E}^*$ and $\mathbf{E}^{**}$ of the nominal exposure variable to be compared, and we need to define these categories in terms of the $k - 1$ dummy variables. Note that we have used *bold letters* to *identify the two categories of E*; this has been done because the E variable is a collection of dummy variables rather than a single variable.

EXAMPLE

E = occupational status (four categories)

$\mathbf{E}^* = \text{category 3 vs. E}^{**} = \text{category 1}$

$\mathbf{E}^* = (\text{OCC}_1^* = 0, \text{OCC}_2^* = 0, \text{OCC}_3^* = 1)$

$\mathbf{E}^{**} = (\text{OCC}_1^{**} = 1, \text{OCC}_2^{**} = 0, \text{OCC}_3^{**} = 0)$

For the occupational status example, suppose we want an odds ratio comparing occupational category 3 with occupational category 1. Here, $\mathbf{E}^*$ represents category 3 and $\mathbf{E}^{**}$ represents category 1. In terms of the three dummy variables for occupational status, then, $\mathbf{E}^*$ is defined by $\text{OCC}_1^* = 0$, $\text{OCC}_2^* = 0$, and $\text{OCC}_3^* = 1$, whereas $\mathbf{E}^{**}$ is defined by $\text{OCC}_1^{**} = 1$, $\text{OCC}_2^{**} = 0$, and $\text{OCC}_3^{**} = 0$.

Generally, define $\mathbf{E}^*$ and $\mathbf{E}^{**}$ as

$$\mathbf{E}^* = (E_1^*, E_2^*, \ldots, E_{k-1}^*)$$

and

$$\mathbf{E}^* = (E_1^{**}, E_2^{**}, \ldots, E_{k-1}^{**})$$

More generally, category $\mathbf{E}^*$ is defined by the dummy variable values E_1^*, E_2^*, and so on up to E_{k-1}^*, which are 0s or 1s. Similarly, category E_1^{**} is defined by the values E_1^{**}, E_2^{**}, and so on up to E_{k-1}^{**}, which is a different specification of 0s or 1s.

No interaction model

$$\begin{aligned}&\text{ROR}_{E^* \text{ vs. } E^{**}}\\ &\quad = \exp[(E_1^* - E_1^{**})\beta_1 + (E_2^* - E_2^{**})\beta_2\\ &\qquad + \ldots + (E_{k-1}^* - E_{k-1}^{**})\beta_{k-1}]\end{aligned}$$

The *general odds ratio formula* for comparing two categories, $\mathbf{E}^*$ vs. $\mathbf{E}^{**}$ of a general nominal exposure variable in a *no interaction logistic model*, is given by the formula ROR equals e to the quantity (E_1^* – E_1^{**}) times β_1 plus (E_2^* – E_2^{**}) times β_2, and so on up to (E_{k-1}^*– E_{k-1}^{**}) times β_{k-1}. When applied to a specific situation, this formula will usually involve more than one β_i in the exponent.

EXAMPLE (OCC)

$$\begin{aligned}\text{ROR}_{3 \text{ vs. } 1} = \exp\Big[&(\overset{0}{\text{OCC}}{}_1^* - \overset{1}{\text{OCC}}{}_1^{**})\beta_1\\ &+ \left(\overset{0}{\text{OCC}}{}_2^* - \overset{0}{\text{OCC}}{}_2^{**}\right)\beta_2\\ &+ \left(\overset{1}{\text{OCC}}{}_3^* - \overset{0}{\text{OCC}}{}_3^{**}\right)\beta_3\Big]\end{aligned}$$

For example, when comparing occupational status category 3 with category 1, the odds ratio formula is computed as e to the quantity (OCC$_1^*$ – OCC$_1^{**}$) times β_1 plus (OCC$_2^*$ – OCC$_2^{**}$) times β_2 plus (OCC$_3^*$ – OCC$_3^{**}$) times β_3.

$$\begin{aligned}&= \exp[(0-1)\beta_1 + (0-0)\beta_2 + (1-0)\beta_3]\\ &= \exp[(-1)\beta_1 + (0)\beta_2 + (1)\beta_3]\\ &= \exp(-\beta_1 + \beta_3)\end{aligned}$$

When we plug in the values for OCC* and OCC**, this expression equals e to the quantity (0 – 1) times β_1 plus (0 – 0) times β_2 plus (1 – 0) times β_3, which equals e to −1 times β_1 plus 0 times β_2 plus 1 times β_3, which reduces to e to the quantity ($-\beta_1$) plus β_3.

$$\widehat{\text{ROR}} = \exp\left(-\hat{\beta}_1 + \hat{\beta}_3\right)$$

We can obtain a single value for the estimate of this odds ratio by fitting the model and replacing β_1 and β_3 with their corresponding estimates $\hat{\beta}_1$ and $\hat{\beta}_3$. Thus, $\widehat{\text{ROR}}$ for this example is given by e to the quantity (– $\hat{\beta}_1$) plus $\hat{\beta}_3$.

$\mathbf{E}^*$ = category 3 vs. E** = category 2:

$\mathbf{E}^*$ = (OCC$_1^*$= 0, OCC$_2^*$= 0, OCC$_3^*$= 1)

$\mathbf{E}^{**}$ = (OCC$_1^{**}$ = 0, OCC$_2^{**}$ = 1, OCC$_3^{**}$ = 0)

In contrast, if category 3 is compared to category 2, then $\mathbf{E}^*$ takes on the values 0, 0, and 1 as before, whereas $\mathbf{E}^{**}$ is now defined by OCC$_1^{**}$ = 0, OCC$_2^{**}$ = 1, and OCC$_3^{**}$ = 0.

$$\begin{aligned}\text{ROR}_{3 \text{ vs. } 2} = &\exp[(0-0)\beta_1 + (0-1)\beta_2\\ &+ (1-0)\beta_3]\\ = &\exp[(0)\beta_1 + (-1)\beta_2 + (1)\beta_3]\\ = &\exp(-\beta_2 + \beta_3)\end{aligned}$$

The odds ratio is then computed as e to the (0 – 0) times β_1 plus (0 – 1) times β_2 plus (1 – 0) times β_3, which equals e to the 0 times β_1 plus −1 times β_2 plus 1 times β_3, which reduces to e to the quantity ($-\beta_2$) plus β_3.

Note. ROR$_{3 \text{ vs. } 1}$ = exp (– β_1 + β_3)

This odds ratio expression involves β_2 and β_3, whereas the previous odds ratio expression that compared category 3 with category 1 involved β_1 and β_3.

V. The Model and Odds Ratio for Several Exposure Variables (No Interaction Case)

q variables: $E_1, E_2, \ldots, E_q$
(dichotomous, ordinal, or interval)

We now consider the odds ratio formula when there are several different exposure variables in the model, rather than a single exposure variable with several categories. The formula for this situation is actually no different than for a single nominal variable. The different exposure variables may be denoted by E_1, E_2, and so on up through E_q. However, rather than being dummy variables, these Es can be any kind of variable – dichotomous, ordinal, or interval.

EXAMPLE

$E_1 = \text{SMK (0,1)}$

$E_2 = \text{PAL (ordinal)}$

$E_3 = \text{SBP (interval)}$

For example, E_1 may be a (0, 1) variable for smoking (SMK), E_2 may be an ordinal variable for physical activity level (PAL), and E_3 may be the interval variable systolic blood pressure (SBP).

No interaction model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1 E_1 + \beta_2 E_2 + \ldots + \beta_q E_q + \sum_{i=1}^{p_1} \gamma_i V_i$$

- $q \neq k - 1$ in general

EXAMPLE

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1 \text{SMK} + \beta_2 \text{PAL} + \beta_3 \text{SBP} + \sum_{i=1}^{p_1} \gamma_i V_i$$

A no interaction model with several exposure variables then takes the form logit P($\mathbf{X}$) equals α plus β_1 times E_1 plus β_2 times E_2, and so on up to β_q times E_q plus the usual set of V terms. This model form is the same as that for a single nominal exposure variable, although this time there are q Es of any type, whereas previously we had $k - 1$ dummy variables to indicate k exposure categories. The corresponding model involving the three exposure variables SMK, PAL, and SBP is shown here.

$\mathbf{E}^*$ vs. $\mathbf{E}^{**}$
$\mathbf{E}^* = (E_1^*, E_2^*, \ldots, E_q^*)$
$\mathbf{E}^{**} = (E_1^{**}, E_2^{**}, \ldots, E_q^{**})$

As before, the general odds ratio formula for several variables requires specifying the values of the exposure variables for two different persons or groups to be compared – denoted by the bold $\mathbf{E}^*$ and $\mathbf{E}^{**}$. Category $\mathbf{E}^*$ is specified by the variable values E_1^*, E_2^*, and so on up to E_q^*, and category $\mathbf{E}^{**}$ is specified by a different collection of values E_1^{**}, E_2^{**}, and so on up to E_q^{**}.

General formula: $E_1, E_2, \ldots, E_8$
(no interaction)

$$\boxed{\begin{aligned} \text{ROR}_{\text{E}^* \text{ vs. E}^{**}} = \exp\Big[&(E_1^* - E_1^{**})\beta_1 \\ &+ (E_2^* - E_2^{**})\beta_2 + \cdots \\ &+ \left(E_q^* - E_q^{**}\right)\beta_q\Big] \end{aligned}}$$

The general odds ratio formula for comparing $\mathbf{E}^*$ vs. $\mathbf{E}^{**}$ is given by the formula ROR equals $\mathbf{e}$ to the quantity $(E_1^* - E_1^*)$ times β_1 plus $(E^* - E^{**})$ times β_2, and so on up to $(E_q^* - E_q^{**})$ times β_q.

In general

- q variables $\neq k - 1$ dummy variables

This formula is the same as that for a single exposure variable with several categories, except that here we have q variables, whereas previously we had $k - 1$ dummy variables.

EXAMPLE

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha &+ \beta_1\text{SMK} + \beta_2\text{PAL}\\ &+ \beta_3\text{SBP}\\ &+ \gamma_1\text{AGE} + \gamma_2\text{SEX}\end{aligned}$$

As an example consider the three exposure variables defined above – SMK, PAL, and SBP. The control variables are AGE and SEX, which are defined in the model as V terms.

Nonsmoker, PAL = 25, SBP = 160
vs.
Smoker, PAL = 10, SBP = 120

$\mathbf{E}^* = (\text{SMK}^* = 0, \text{PAL}^* = 25, \text{SBP}^* = 160)$

$\mathbf{E}^{**} = (\text{SMK}^{**} = 1, \text{PAL}^{**} = 10, \text{SBP}^{**} = 120)$

Suppose we wish to compare a nonsmoker who has a PAL score of 25 and systolic blood pressure of 160 to a smoker who has a PAL score of 10 and systolic blood pressure of 120, controlling for AGE and SEX. Then, here, $\mathbf{E}^*$ is defined by $\text{SMK}^* = 0$, $\text{PAL}^* = 25$, and $\text{SBP}^* = 160$, whereas $\mathbf{E}^{**}$ is defined by $\text{SMK}^{**} = 1$, $\text{PAL}^{**} = 10$, and $\text{SBP}^{**} = 120$.

AGE and SEX fixed, but unspecified

The control variables AGE and SEX are considered fixed but do not need to be specified to obtain an odds ratio because the model contains no interaction terms.

$$\begin{aligned}\text{ROR}_{\mathbf{E}^*\text{ vs. }\mathbf{E}^{**}} = \exp\big[&(\text{SMK}^* - \text{SMK}^{**})\beta_1\\ &+ (\text{PAL}^* - \text{PAL}^{**})\beta_2\\ &+ (\text{SBP}^* - \text{SBP}^{**})\beta_3\big]\end{aligned}$$

The odds ratio is then computed as e to the quantity $(\text{SMK}^* - \text{SMK}^{**})$ times β_1 plus $(\text{PAL}^* - \text{PAL}^{**})$ times β_2 plus $(\text{SBP}^* - \text{SBP}^{**})$ times β_3,

$$\begin{aligned}= \exp[&(0 - 1)\beta_1 + (25 - 10)\beta_2\\ &+ (160 - 120)\beta_3]\end{aligned}$$

which equals e to $(0 - 1)$ times β_1 plus $(25 - 10)$ times β_2 plus $(160 - 120)$ times β_3,

$$\begin{aligned}= \exp[&(-1)\beta_1 + (15)\beta_2\\ &+ (40)\beta_3]\end{aligned}$$

which equals e to the quantity $- 1$ times β_1 plus 15 times β_2 plus 40 times β_3,

$$= \exp(-\beta_1 + 15\beta_2 + 40\beta_3)$$

which reduces to e to the quantity $- \beta_1$ plus $15\beta_2$ plus $40\beta_3$.

$$\widehat{\text{ROR}} = \exp\left(-\hat{\beta}_1 + 15\hat{\beta}_2 + 40\hat{\beta}_3\right)$$

An estimate of this odds ratio can then be obtained by fitting the model and replacing β_1, β_2, and β_3 by their corresponding estimates $\hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$. Thus, $\widehat{\text{ROR}}$ equals e to the quantity $- \hat{\beta}_1$ plus $15\hat{\beta}_2$ plus $40\hat{\beta}_3$.

ANOTHER EXAMPLE

$\mathbf{E}^* = (\text{SMK}^* = 1, \text{PAL}^* = 25, \text{SBP}^* = 160)$

$\mathbf{E}^{**} = (\text{SMK}^{**} = 1, \text{PAL}^{**} = 5, \text{SBP}^{**} = 200)$

controlling for AGE and SEX

$$\begin{aligned}\text{ROR}_{\mathbf{E}^*\text{ vs. }\mathbf{E}^{**}} &= \exp[(1-1)\beta_1+(25-5)\beta_2 \\ &\quad +(160-200)\beta_3] \\ &= \exp[(0)\beta_1+(20)\beta_2 \\ &\quad +(-40)\beta_3] \\ &= \exp(20\beta_2-40\beta_3)\end{aligned}$$

As a second example, suppose we compare a smoker who has a PAL score of 25 and a systolic blood pressure of 160 to a smoker who has a PAL score of 5 and a systolic blood pressure of 200, again controlling for AGE and SEX.

The ROR is then computed as e to the quantity (1 – 1) times β_1 plus (25 – 5) times β_2 plus (160 – 200) times β_3, which equals e to 0 times β_1 plus 20 times β_2 plus –40 times β_3, which reduces to e to the quantity $20\beta_2$ minus $40\beta_3$.

VI. The Model and Odds Ratio for Several Exposure Variables with Confounders and Interaction

We now consider a final situation involving *several exposure variables, confounders* (i.e., *V*s), and *interaction variables* (i.e., *W*s), where the *W*s go into the model as product terms with one of the *E*s.

EXAMPLE: The Variables

$E_1 = \text{SMK}, E_2 = \text{PAL}, E_3 = \text{SBP}$

$V_1 = \text{AGE} = W_1, V_2 = \text{SEX} = W_2$

$E_1W_1 = \text{SMK} \times \text{AGE}, E_1W_2 = \text{SMK} \times \text{SEX}$

$E_2W_1 = \text{PAL} \times \text{AGE}, E_2W_2 = \text{PAL} \times \text{SEX}$

$E_3W_1 = \text{SBP} \times \text{AGE}, E_3W_2 = \text{SBP} \times \text{SEX}$

As an example, we again consider the three exposures SMK, PAL, and SBP and the two control variables AGE and SEX. We add to this list product terms involving each exposure with each control variable. These product terms are shown here.

EXAMPLE: The Model

$$\begin{aligned}\text{logit P}(\mathbf{X}) &= \alpha+\beta_1\text{SMK}+\beta_2\text{PAL} \\ &\quad +\beta_3\text{SBP}+\gamma_1\text{AGE}+\gamma_2\text{SEX} \\ &\quad +\text{SMK}(\delta_{11}\text{AGE}+\delta_{12}\text{SEX}) \\ &\quad +\text{PAL}(\delta_{21}\text{AGE}+\delta_{22}\text{SEX}) \\ &\quad +\text{SBP}(\delta_{31}\text{AGE}+\delta_{32}\text{SEX})\end{aligned}$$

The corresponding model is given by logit P(**X**) equals α plus β_1 times SMK plus β_2 times PAL plus β_3 times SBP plus the sum of *V* terms involving AGE and SEX plus SMK times the sum of δ times *W* terms, where the *W*s are AGE and SEX, plus PAL times the sum of additional δ times *W* terms, plus SBP times the sum of additional δ times *W* terms. Here the δs are coefficients of interaction terms involving one of the three exposure variables – either SMK, PAL, or SEX – and one of the two control variables – either AGE or SEX.

EXAMPLE: The Odds Ratio

$\mathbf{E}^*$ vs. $\mathbf{E}^{**}$

$\mathbf{E}^* = (\text{SMK}^* = 0, \text{PAL}^* = 25, \text{SBP}^* = 160)$

$\mathbf{E}^{**} = (\text{SMK}^{**} = 1, \text{PAL}^{**} = 10, \text{SBP}^{**} = 120)$

To obtain an odds ratio expression for this model, we again must identify two specifications of the collection of exposure variables to be compared. We have referred to these specifications generally by the bold terms $\mathbf{E}^*$ and $\mathbf{E}^{**}$. In the above example, $\mathbf{E}^*$ is defined by $\text{SMK}^* = 0$, $\text{PAL}^* = 25$, and $\text{SBP}^* = 160$, whereas $\mathbf{E}^{**}$ is defined by $\text{SMK}^{**} = 1$, $\text{PAL}^{**} = 10$, and $\text{SBP}^{**} = 120$.

ROR (no interaction): βs only

ROR (interaction): βs and δs

The previous odds ratio formula that we gave for several exposures but no interaction involved only β coefficients for the exposure variables. Because the model we are now considering contains interaction terms, the corresponding odds ratio will involve not only the β coefficients, but also δ coefficients for all interaction terms involving one or more exposure variables.

EXAMPLE (continued)

$$
\begin{aligned}
\text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} = \exp[&(\text{SMK}^* - \text{SMK}^{**})\beta_1 \\
&+ (\text{PAL}^* - \text{PAL}^{**})\beta_2 \\
&+ (\text{SBP}^* - \text{SBP}^{**})\beta_3 \\
&+ \delta_{11}(\text{SMK}^* - \text{SMK}^{**})\text{AGE} \\
&+ \delta_{12}(\text{SMK}^* - \text{SMK}^{**})\text{SEX} \\
&+ \delta_{21}(\text{PAL}^* - \text{PAL}^{**})\text{AGE} \\
&+ \delta_{22}(\text{PAL}^* - \text{PAL}^{**})\text{SEX} \\
&+ \delta_{31}(\text{SBP}^* - \text{SBP}^{**})\text{AGE} \\
&+ \delta_{32}(\text{SBP}^* - \text{SBP}^{**})\text{SEX}
\end{aligned}
$$

The odds ratio formula for our example then becomes e to the quantity (SMK* – SMK**) times β_1 plus (PAL* – PAL**) times β_2 plus (SBP* – SBP**) times β_3 plus the sum of terms involving a δ coefficient times the difference between E^* and E^{**} values of one of the exposures times a W variable.

For example, the first of the interaction terms is δ_{11} times the difference (SMK* – SMK**) times AGE, and the second of these terms is δ_{12} times the difference (SMK* – SMK**) times SEX.

$$
\begin{aligned}
\text{ROR} = \exp[&(0 - 1)\beta_1 + (25 - 10)\beta_2 \\
&+ (160 - 120)\beta_3
\end{aligned}
$$

interaction with SMK

$$+ \delta_{11}\,(0 - 1)\text{ AGE} + \delta_{12}\,(0 - 1)\text{ SEX}$$

interaction with PAL

$$+ \delta_{21}\,(25 - 10)\text{ AGE} + \delta_{22}\,(25 - 10)\text{ SEX}$$

interaction with SBP

$$+ \delta_{31}\,(160 - 120)\text{ AGE} + \delta_{32}\,(160 - 120)\text{ SEX}$$

When we substitute into the odds ratio formula the values for $\mathbf{E}^*$ and $\mathbf{E}^{**}$, we obtain the expression e to the quantity (0 – 1) times β_1 plus (25 – 10) times β_2 plus (160 – 120) times β_3 plus several terms involving interaction coefficients denoted as δs.

The first set of these terms involves interactions of AGE and SEX with SMK. These terms are δ_{11} times the difference (0 – 1) times AGE plus δ_{12} times the difference (0 – 1) times SEX. The next set of δ terms involves interactions of AGE and SEX with PAL. The last set of δ terms involves interactions of AGE and SEX with SBP.

$$
\begin{aligned}
= \exp(&-\beta_1 + 15\beta_2 + 40\beta_3 \\
&- \delta_{11}\text{AGE} - \delta_{12}\text{SEX} \\
&+ 15\delta_{21}\text{AGE} + 15\delta_{22}\text{SEX} \\
&+ 40\delta_{31}\text{AGE} + 40\delta_{32}\text{SEX})
\end{aligned}
$$

After subtraction, this expression reduces to the expression shown here at the left.

$$
\begin{aligned}
= \exp(&-\beta_1 + 15\beta_2 + 40\beta_3 \\
&+ \text{AGE}(-\delta_{11} + 15\delta_{21} + 40\delta_{31}) \\
&+ \text{SEX}(-\delta_{12} + 15\delta_{22} + 40\delta_{32})]
\end{aligned}
$$

We can simplify this expression further by factoring out AGE and SEX to obtain e to the quantity minus β_1 plus 15 times β_2 plus 40 times β_3 plus AGE times the quantity minus δ_{11} plus 15 times δ_{21} plus 40 times δ_{31} plus SEX times the quantity minus δ_{12} plus 15 times δ_{22} plus 40 times δ_{32}.

EXAMPLE (continued)

Note. Specify AGE and SEX to get a numerical value.

e.g., AGE = 35, SEX = 1:

$$\widehat{\text{ROR}} = \exp[-\hat{\beta}_1 + 15\hat{\beta}_2 + 40\hat{\beta}_3$$

$$\text{AGE} \rightarrow + 35(-\hat{\delta}_{11} + 15\hat{\delta}_{21} + 40\hat{\delta}_{31})$$

$$\text{SEX} \rightarrow + 1(-\hat{\delta}_{12} + 15\hat{\delta}_{22} + 40\hat{\delta}_{32})]$$

$$\begin{aligned}\widehat{\text{ROR}} = \exp\big(&-\hat{\beta}_1 + 15\hat{\beta}_2 + 40\hat{\beta}_3\\ &- 35\hat{\delta}_{11} + 525\hat{\delta}_{21} + 1400\hat{\delta}_{31}\\ &- \hat{\delta}_{12} + 15\hat{\delta}_{22} + 40\hat{\delta}_{32}\big)\end{aligned}$$

Note that this expression tells us that once we have fitted the model to the data to obtain estimates of the β and δ coefficients, we must specify values for the effect modifiers AGE and SEX before we can get a numerical value for the odds ratio. In other words, the odds ratio will give a different numerical value depending on which values we specify for the effect modifiers AGE and SEX.

For instance, if we choose AGE equals 35 and SEX equals 1 say, for females, then the estimated odds ratio becomes the expression shown here.

This odds ratio expression can alternatively be written as e to the quantity minus $\hat{\beta}_1$ plus 15 times $\hat{\beta}_2$ plus 40 times $\hat{\beta}_3$ minus 35 times $\hat{\delta}_{11}$ plus 525 times $\hat{\delta}_{21}$ plus 1,400 times $\hat{\delta}_{31}$ minus $\hat{\delta}_{12}$ plus 15 times $\hat{\delta}_{22}$ plus 40 times $\hat{\delta}_{32}$. This expression will give us a single numerical value for 35-year-old females once the model is fitted and estimated coefficients are obtained.

General model
- Several exposures
- Confounders
- Effect modifiers

We have just worked through a specific example of the odds ratio formula for a model involving several exposure variables and controlling for both confounders and effect modifiers. To obtain a general odds ratio formula for this situation, we first need to write the model in general form.

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha + \beta_1 E_1 + \beta_2 E_2 + \cdots\\ + \beta_q E_q + \sum_{i=1}^{p_1} \gamma_i V_i\\ + E_1 \sum_{j=1}^{p_2} \delta_{1j} W_i\\ + E_2 \sum_{j=1}^{p_2} \delta_{2j} W_j + \cdots\\ + E_q \sum_{j=1}^{p_2} \delta_{qj} W_j\end{aligned}$$

This expression is given by the logit of P($\mathbf{X}$) equals α plus β_1 times E_1 plus β_2 times E_2, and so on up to β_q times E_q plus the usual set of V terms of the form $\gamma_i V_i$ plus the sum of additional terms, each having the form of an exposure variable times the sum of δ times W terms. The first of these interaction expressions is given by E_1 times the sum of δ_{1j} times W_j, where E_1 is the first exposure variable, δ_{1j} is an unknown coefficient, and W_j is the jth effect modifying variable. The last of these terms is E_q times the sum of δ_{qj} times W_j, where E_q is the last exposure variable, δ_{qj} is an unknown coefficient, and W_j is the jth effect modifying variable.

We assume the same W_j for each exposure variable

e.g., AGE and SEX are *W*s for each *E*.

Note that this model assumes that the same effect modifying variables are being considered for each exposure variable in the model, as illustrated in our preceding example above with AGE and SEX.

A more general model can be written that allows for different effect modifiers corresponding to different exposure variables, but for convenience, we limit our discussion to a model with the same modifiers for each exposure variable.

Odds ratio for several *E*s:

$$\mathbf{E}^* = \left(E_1^*, E_2^*, \ldots, E_q^*\right)$$

$$\mathbf{E}^{**} = \left(E_1^{**}, E_2^{**}, \ldots, E_q^{**}\right)$$

To obtain an odds ratio expression for the above model involving several exposures, confounders, and interaction terms, we again must identify two specifications of the exposure variables to be compared. We have referred to these specifications generally by the bold terms $\mathbf{E}^*$ and $\mathbf{E}^{**}$. Group $\mathbf{E}^*$ is specified by the variable values E_1^*, E_2^*, and so on up to E_q^*; group $\mathbf{E}^{**}$ is specified by a different collection of values E_1^{**}, E_2^{**}, and so on up to E_q^{**}.

General Odds Ratio Formula:

$$\boxed{\begin{aligned}\mathrm{ROR}_{E^*\text{ vs. }E^{**}} = \exp\Bigg[&(E_1^* - E_1^{**})\beta_1 \\ &+ (E_2^* - E_2^{**})\beta_2 \\ &+ \cdots + \left(E_q^* - E_q^{**}\right)\beta_q \\ &+ (E_1^* - E_1^{**})\sum_{j=1}^{p_2}\delta_{1j}W_j \\ &+ (E_2^* - E_2^{**})\sum_{j=1}^{p_2}\delta_{2j}W_j \\ &+ \cdots \\ &+ \left(E_q^* - E_q^{**}\right)\times\sum_{j=1}^{p_2}\delta_{\mathrm{q}j}W_j\Bigg]\end{aligned}}$$

The general odds ratio formula for comparing two such specifications, $\mathbf{E}^*$ vs. $\mathbf{E}^{**}$, is given by the formula ROR equals e to the quantity ($E_1^* - E_1^{**}$) times β_1 plus ($E_2^* - E_2^{**}$) times β_2, and so on up to ($E_q^* - E_q^{**}$) times β_q plus the sum of terms of the form ($\mathbf{E}^* - \mathbf{E}^{**}$) times the sum of δ times *W*, where each of these latter terms correspond to interactions involving a different exposure variable.

EXAMPLE: $q = 3$

$$
\begin{aligned}
\mathrm{ROR}_{\mathbf{E}^{*}\,\mathrm{vs.}\,\mathbf{E}^{**}} = \exp\big[&(\mathrm{SMK}^{*}-\mathrm{SMK}^{**})\beta_1\\
&+(\mathrm{PAL}^{*}-\mathrm{PAL}^{**})\beta_2\\
&+(\mathrm{SBP}^{*}-\mathrm{SBP}^{**})\beta_3\\
&+\delta_{11}(\mathrm{SMK}^{*}-\mathrm{SMK}^{**})\mathrm{AGE}\\
&+\delta_{12}(\mathrm{SMK}^{*}-\mathrm{SMK}^{**})\mathrm{SEX}\\
&+\delta_{21}(\mathrm{PAL}^{*}-\mathrm{PAL}^{**})\mathrm{AGE}\\
&+\delta_{22}(\mathrm{PAL}^{*}-\mathrm{PAL}^{**})\mathrm{SEX}\\
&+\delta_{31}(\mathrm{SBP}^{*}-\mathrm{SBP}^{**})\mathrm{AGE}\\
&+\delta_{32}(\mathrm{SBP}^{*}-\mathrm{SBP}^{**})\mathrm{SEX}\big]
\end{aligned}
$$

In our previous example using this formula, there are q equals three exposure variables (namely, SMK, PAL, and SBP), two confounders (namely, AGE and SEX), which are in the model as V variables, and two effect modifiers (also AGE and SEX), which are in the model as W variables. The odds ratio expression for this example is shown here again.

This odds ratio expression does not contain coefficients for the confounding effects of AGE and SEX. Nevertheless, these effects are being controlled because AGE and SEX are contained in the model as V variables in addition to being W variables.

- AGE and SEX controlled as *V*s as well as *W*s
- RORs depend on values of *W*s (AGE and SEX)

Note that for this example, as for any model containing interaction terms, the odds ratio expression will yield different values for the odds ratio depending on the values of the effect modifiers – in this case, AGE and SEX – that are specified.

SUMMARY

Chapters up to this point:

1. Introduction
2. Important Special Cases

✓ 3. Computing the Odds Ratio

This presentation is now complete. We have described how to compute the odds ratio for an arbitrarily coded single exposure variable that may be dichotomous, ordinal, or interval. We have also described the odds ratio formula when the exposure variable is a polytomous nominal variable like occupational status. And, finally, we have described the odds ratio formula when there are several exposure variables, controlling for confounders without interaction terms and controlling for confounders together with interaction terms.

4. Maximum Likelihood (ML) Techniques: An Overview
5. Statistical Inferences Using ML Techniques

In the next chapter (Chap. 4), we consider how the method of maximum likelihood is used to estimate the parameters of the logistic model. And in Chap. 5, we describe statistical inferences using ML techniques.

Detailed Outline

I. Overview (pages 76–77)

A. Focus: computing OR for *E, D* relationship adjusting for confounding and effect modification.

B. Review of the special case – the *E, V, W* model:

i. The model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \sum_{i=1}^{p_1} \gamma_i V_i + E \sum_{j=1}^{p_2} \delta_j W_j.$$

ii. Odds ratio formula for the *E, V, W* model, where *E* is a (0, 1) variable:

$$\text{ROR}_{E=1 \text{ vs. } E=0} = \exp\left(\beta + \sum_{j=1}^{p_2} \delta_j W_j\right).$$

II. Odds ratio for other codings of a dichotomous *E* (pages 77–79)

A. For the *E, V, W* model with *E* coded as $E = a$ if exposed and as $E = b$ if unexposed, the odds ratio formula becomes

$$\text{ROR}_{E=1 \text{ vs. } E=0} = \exp\left[(a-b)\beta + (a-b)\sum_{j=1}^{p_2} \delta_j W_j\right]$$

B. Examples: $a = 1,\quad b = 0$: $\text{ROR} = \exp(\beta)$
$a = 1,\quad b = -1$: $\text{ROR} = \exp(2\beta)$
$a = 100, b = 0$: $\text{ROR} = \exp(100\beta)$

C. Final computed odds ratio has the same value provided the correct formula is used for the corresponding coding scheme, even though the coefficients change as the coding changes.

D. Numerical example from Evans County study.

III. Odds ratio for arbitrary coding of *E* (pages 79–82)

A. For the *E, V, W* model where $\mathbf{E}^*$ and $\mathbf{E}^{**}$ are any two values of *E* to be compared, the odds ratio formula becomes

$$\text{ROR}_{E^* \text{ vs. } E^{**}} = \exp\left[\left(E^* - E^{**}\right)\beta + \left(E^* - E^{**}\right)\sum_{j=1}^{p_2} \delta_j W_j\right]$$

B. Examples: $E = \text{SSU} =$ social support status (0–5)
$E = \text{SBP} =$ systolic blood pressure (interval).

C. No interaction odds ratio formula:

$$\text{ROR}_{E^* \text{ vs. } E^{**}} = \exp\left[\left(E^* - E^{**}\right)\beta\right].$$

D. Interval variables, e.g., SBP: Choose values for comparison that represent clinically meaningful categories, e.g., quintiles.

IV. The model and odds ratio for a nominal exposure variable (no interaction case) (pages 82–84)

A. No interaction model involving a nominal exposure variable with k categories:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1 E_1 + \beta_2 E_2 + \cdots + \beta_{k-1} E_{k-1} + \sum_{i=1}^{p_1} \gamma_i V_i,$$

where $E_1, E_2, \ldots, E_{k-1}$ denote $k - 1$ dummy variables that distinguish the k categories of the nominal exposure variable denoted as $\mathbf{E}$, i.e., $E_i = 1$ if category i or 0 if otherwise.

B. Example of model involving $k = 4$ categories of occupational status:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1 \text{OCC}_1 + \beta_2 \text{OCC}_2 + \beta_3 \text{OCC}_3 + \sum_{i=1}^{p_1} \gamma_i V_i,$$

where OCC_1, OCC_2, and OCC_3 denote $k - 1 = 3$ dummy variables that distinguish the four categories of occupation.

C. Odds ratio formula for no interaction model involving a nominal exposure variable:

$$\text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} = \exp\left[\begin{array}{l} (E_1^* - E_1^{**})\beta_1 + (E_2^* - E_2^{**})\beta_2 \\ + \cdots + (E_{k-1}^* - E_{k-1}^{**})\beta_{k-1} \end{array}\right],$$

where $\mathbf{E}^* = (E_1^*, E_2^*, \ldots, E_{k-1}^*)$ and $\mathbf{E}^{**} = (E_1^{**}, E_2^{**}, \ldots, E_{k-\,1}^{**})$ are two specifications of the set of dummy variables for $\mathbf{E}$ to be compared.

D. Example of odds ratio involving $k = 4$ categories of occupational status:

$$\text{ROR}_{\text{OCC}^* \text{ vs. OCC}^{**}} = \exp\left[\begin{array}{l} (\text{OCC}_1^* - \text{OCC}_1^{**})\beta_1 + (\text{OCC}_2^* - \text{OCC}_2^{**})\beta_2 \\ + (\text{OCC}_3^* - \text{OCC}_3^{**})\beta_3 \end{array}\right].$$

V. The model and odds ratio for several exposure variables (no interaction case) (pages 85–87)

A. The model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1 E_1 + \beta_2 E_2 + \cdots + \beta_q E_q + \sum_{i=1}^{p_1} \gamma_i V_i,$$

where $E_1, E_2, \ldots, E_q$ denote q exposure variables of interest.

B. Example of model involving three exposure variables:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1\text{SMK} + \beta_2\text{PAL} + \beta_3\text{SBP} + \sum_{i=1}^{p_1} \gamma_i V_i.$$

C. The odds ratio formula for the general no interaction model:

$$\text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} = \exp[(E_1^* - E_1^{**})\beta_1 + (E_2^* - E_2^{**})\beta_2 + \cdots + (E_q^* - E_q^{**})\beta_q],$$

where $\mathbf{E}^* = (E_1^*, E_2^*, \ldots, E_q^*)$ and $\mathbf{E}^{**} = (E_1^*, E_2^{**}, \ldots, E_q^{**})$ are two specifications of the collection of exposure variables to be compared.

D. Example of odds ratio involving three exposure variables:

$$\text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} = \exp[(\text{SMK}^* - \text{SMK}^{**})\beta_1 + (\text{PAL}^* - \text{PAL}^{**})\beta_2 + (\text{SBP}^* - \text{SBP}^{**})\beta_3].$$

VI. The model and odds ratio for several exposure variables with confounders and interaction (pages 87–91)

A. An example of a model with three exposure variables:

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha &+ \beta_1\text{SMK} + \beta_2\text{PAL} + \beta_3\text{SBP} + \gamma_1\text{AGE} \\ &+ \gamma_2\text{SEX} + \text{SMK}(\delta_{11}\text{AGE} + \delta_{12}\text{SEX}) \\ &+ \text{PAL}(\delta_{21}\text{AGE} + \delta_{22}\text{SEX}) \\ &+ \text{SBP}(\delta_{31}\text{AGE} + \delta_{32}\text{SEX}).\end{aligned}$$

B. The odds ratio formula for the above model:

$$\begin{aligned}\text{ROR}_{\mathbf{E}^* \text{vs.} \mathbf{E}^{**}} = \exp[&(\text{SMK}^* - \text{SMK}^{**})\beta_1 \\ &+ (\text{PAL}^* - \text{PAL}^{**})\beta_2 + (\text{SBP}^* - \text{SBP}^{**})\beta_3 \\ &+ \delta_{11}(\text{SMK}^* - \text{SMK}^{**})\text{AGE} \\ &+ \delta_{12}(\text{SMK}^* - \text{SMK}^{**})\text{SEX} \\ &+ \delta_{21}(\text{PAL}^* - \text{PAL}^{**})\text{AGE} \\ &+ \delta_{22}(\text{PAL}^* - \text{PAL}^{**})\text{SEX} \\ &+ \delta_{31}(\text{SBP}^* - \text{SBP}^{**})\text{AGE} \\ &+ \delta_{32}(\text{SBP}^* - \text{SBP}^{**})\text{SEX}]\end{aligned}$$

C. The general model:

$$\begin{aligned}\text{logit } \mathrm{P}(\mathbf{X}) = {} & \alpha + \beta_1 E_1 + \beta_2 E_2 + \cdots + \beta_q E_q \\ & + \sum_{i=1}^{p_1} \gamma_i V_i + E_1 \sum_{j=1}^{p_2} \delta_{1j} W_j \\ & + E_2 \sum_{j=1}^{p_2} \delta_{2j} W_j + \cdots + E_q \sum_{j=1}^{p_2} \delta_{qj} W_j\end{aligned}$$

D. The general odds ratio formula:

$$\begin{aligned}\text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} = \exp\Bigg[& (E_1^* - E_1^{**})\beta_1 + (E_2^* - E_2^{**})\beta_2 \\ & + \cdots + (E_q^* - E_q^{**})\beta_q \\ & + (E_1^* - E_1^{**}) \sum_{j=1}^{p_2} \delta_{1j} W_j \\ & + (E_2^* - E_2^{**}) \sum_{j=1}^{p_2} \delta_{2j} W_j \\ & + \cdots + (E_q^* - E_q^{**}) \sum_{j=1}^{p_2} \delta_{qj} W_j \Bigg]\end{aligned}$$

Practice Exercises

Given the model

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \gamma_1(\text{SMK}) + \gamma_2(\text{HPT}) + \delta_1(E \times \text{SMK}) + \delta_2(E + \text{HPT}),$$

where SMK (smoking status) and HPT (hypertension status) are dichotomous variables.

Answer the following true or false questions (circle T or F):

T F 1. If E is coded as (0 = unexposed, 1 = exposed), then the odds ratio for the E, D relationship that controls for SMK and HPT is given by
$\exp[\beta + \delta_1(E \times \text{SMK}) + \delta_2(E \times \text{HPT})]$.

T F 2. If E is coded as $(-1, 1)$, then the odds ratio for the E, D relationship that controls for SMK and HPT is given by
$\exp[2\beta + 2\delta_1(\text{SMK}) + 2\delta_2(\text{HPT})]$.

T F 3. If there is no interaction in the above model and E is coded as $(-1, 1)$, then the odds ratio for the E, D relationship that controls for SMK and HPT is given by $\exp(\beta)$.

T F 4. If the correct odds ratio formula for a given coding scheme for E is used, then the estimated odds ratio will be the same regardless of the coding scheme used.

Given the model

$$\text{logit P}(\mathbf{X}) = \alpha + \beta(\text{CHL}) + \gamma(\text{AGE}) + \delta(\text{AGE} \times \text{CHL}),$$

where CHL and AGE are continuous variables,

Answer the following true or false questions (circle T or F):

T F 5. The odds ratio that compares a person with CHL = 200 to a person with CHL = 140 controlling for AGE is given by $\exp(60\beta)$.

T F 6. If we assume no interaction in the above model, the expression $\exp(\beta)$ gives the odds ratio for describing the effect of one unit change in CHL value, controlling for AGE.

Suppose a study is undertaken to compare the lung cancer risks for samples from three regions (urban, suburban, and rural) in a certain state, controlling for the potential confounding and effect-modifying effects of AGE, smoking status (SMK), RACE, and SEX.

7. State the logit form of a logistic model that treats region as a polytomous exposure variable and controls for the confounding effects of AGE, SMK, RACE, and SEX. (Assume no interaction involving any covariates with exposure.)
8. For the model of Exercise 7, give an expression for the odds ratio for the *E, D* relationship that compares urban with rural persons, controlling for the four covariates.
9. Revise your model of Exercise 7 to allow effect modification of each covariate with the exposure variable. State the logit form of this revised model.
10. For the model of Exercise 9, give an expression for the odds ratio for the *E, D* relationship that compares urban with rural persons, controlling for the confounding and effect-modifying effects of the four covariates.
11. Given the model

$$\begin{aligned} \text{logit P}(\mathbf{X}) = \alpha + \beta_1(\text{SMK}) + \beta_1(\text{ASB}) + \gamma_1(\text{AGE}) \\ + \delta_1(\text{SMK} \times \text{AGE}) + \delta_2(\text{ASB} \times \text{AGE}), \end{aligned}$$

where SMK is a (0, 1) variable for smoking status, ASB is a (0, 1) variable for asbestos exposure status, and AGE is treated continuously,

Circle the (one) correct choice among the following statements:

a. The odds ratio that compares a smoker exposed to asbestos to a nonsmoker not exposed to asbestos, controlling for age, is given by $\exp(\beta_1 + \beta_2 + \delta_1 + \delta_2)$.
b. The odds ratio that compares a nonsmoker exposed to asbestos to a nonsmoker unexposed to asbestos, controlling for age, is given by $\exp[\beta_2 + \delta_2(\text{AGE})]$.
c. The odds ratio that compares a smoker exposed to asbestos to a smoker unexposed to asbestos, controlling for age, is given by $\exp[\beta_1 + \delta_1(\text{AGE})]$.
d. The odds ratio that compares a smoker exposed to asbestos to a nonsmoker exposed to asbestos, controlling for age, is given by $\exp[\beta_1 + \delta_1(\text{AGE}) + \delta_2(\text{AGE})]$.
e. None of the above statements is correct.

Test

1. Given the following logistic model
   logit P(**X**) = $\alpha + \beta\text{CAT} + \gamma_1\text{AGE} + \gamma_2\text{CHL}$,
   where CAT is a dichotomous exposure variable and AGE and CHL are continuous, *answer the following questions concerning the odds ratio that compares exposed to unexposed persons controlling for the effects of AGE and CHL*:
   a. Give an expression for the odds ratio for the *E, D* relationship, assuming that CAT is coded as (0 = low CAT, 1 = high CAT).
   b. Give an expression for the odds ratio, assuming CAT is coded as (0, 5).
   c. Give an expression for the odds ratio, assuming that CAT is coded as (-1, 1).
   d. Assuming that the same dataset is used for computing odds ratios described in parts a–c above, what is the relationship among odds ratios computed by using the three different coding schemes of parts a–c?
   e. Assuming the same data set as in part d above, what is the relationship between the βs that are computed from the three different coding schemes?
2. Suppose the model in Question 1 is revised as follows:
   logit P(**X**) = $\alpha + \beta\text{CAT} + \gamma_1\text{AGE} + \gamma_2\text{CHL} + \text{CAT}(\delta_1\text{AGE}$
   $+ \delta_2\text{CHL})$.

For this revised model, answer the same questions as given in parts a–e of Question 1.

   a.

   b.

   c.

   d.

   e.

3. Given the model
   logit P(**X**) = $\alpha + \beta\text{SSU} + \gamma_1\text{AGE} + \gamma_2\text{SEX}$
   $+ \text{SSU}(\delta_1\text{AGE} + \delta_2\text{SEX})$,
   where SSU denotes "social support score" and is an ordinal variable ranging from 0 to 5, *answer the following questions about the above model*:

a. Give an expression for the odds ratio that compares a person who has SSU = 5 to a person who has SSU = 0, controlling for AGE and SEX.

b. Give an expression for the odds ratio that compares a person who has SSU = 1 to a person who has SSU = 0, controlling for AGE and SEX.

c. Give an expression for the odds ratio that compares a person who has SSU = 2 to a person who has SSU = 1, controlling for AGE and SEX.

d. Assuming that the same data set is used for parts b and c, what is the relationship between the odds ratios computed in parts b and c?

4. Suppose the variable SSU in Question 3 is partitioned into three categories denoted as *low, medium*, and *high*.

   a. Revise the model of Question 3 to give the logit form of a logistic model that treats SSU as a nominal variable with three categories (*assume no interaction*).

   b. Using your model of part a, give an expression for the odds ratio that compares high to low SSU persons, controlling for AGE and SEX.

   c. Revise your model of part a to allow for effect modification of SSU with AGE and with SEX.

   d. Revise your odds ratio of part b to correspond to your model of part c.

5. Given the following model

   $$\text{logit } P(\mathbf{X}) = \alpha + \beta_1\text{NS} + \beta_2\text{OC} + \beta_3\text{AFS} + \gamma_1\text{AGE} + \gamma_2\text{RACE},$$

   where NS denotes number of sex partners in one's lifetime, OC denotes oral contraceptive use (yes/no), and AFS denotes age at first sexual intercourse experience, *answer the following questions about the above model*:

   a. Give an expression for the odds ratio that compares a person who has NS = 5, OC = 1, and AFS = 26 to a person who has NS = 5, OC = 1, and AFS = 16, controlling for AGE and RACE.

   b. Give an expression for the odds ratio that compares a person who has NS = 200, OC = 1, and AFS = 26 to a person who has NS = 5, OC = 1, and AFS = 16, controlling for AGE and RACE.

6. Suppose the model in Question 5 is revised to contain interaction terms:

$$\begin{aligned}\text{logit P}(\mathbf{X}) = {} & \alpha+\beta_1\text{NS}+\beta_2\text{OC}+\beta_3\text{AFS}+\gamma_1\text{AGE}+\gamma_2\text{RACE}\\ & +\delta_{11}(\text{NS}\times\text{AGE})+\delta_{12}(\text{NS}\times\text{RACE})\\ & +\delta_{21}(\text{OC}\times\text{AGE})+\delta_{22}(\text{OC}\times\text{RACE})\\ & +\delta_{31}(\text{AFS}\times\text{AGE})+\delta_{32}(\text{AFS}\times\text{RACE}).\end{aligned}$$

For this revised model, answer the same questions as given in parts a and b of Question 5.

a.

b.

Answers to Practice Exercises

1. F: the correct odds ratio expression is $\exp[\beta + \delta_1(\text{SMK}) + \delta_2(\text{HPT})]$
2. T
3. F: the correct odds ratio expression is $\exp(2\beta)$
4. T
5. F: the correct odds ratio expression is $\exp[60\beta + 60\delta(\text{AGE})]$
6. T
7. $$\text{logit } P(\mathbf{X}) = \alpha + \beta_1 R_1 + \beta_2 R_2 + \gamma_1\text{AGE} + \gamma_2\text{SMK} + \gamma_3\text{RACE} + \gamma_4\text{SEX},$$
   where R_1 and R_2 are dummy variables indicating region, e.g., $R_1 = (1$ if urban, 0 if other) and $R_2 = (1$ if suburban, 0 if other).
8. When the above coding for the two dummy variables is used, the odds ratio that compares urban with rural persons is given by $\exp(\beta_1)$.
9. $$\begin{aligned}\text{logit } P(\mathbf{X}) = \alpha &+ \beta_1 R_1 + \beta_2 R_2 + \gamma_1\text{AGE} + \gamma_2\text{SMK} \\ &+ \gamma_3\text{RACE} + \gamma_4\text{SEX} + R_1(\delta_{11}\text{AGE} + \delta_{12}\text{SMK} \\ &+ \delta_{13}\text{RACE} + \delta_{14}\text{SEX}) + R_2(\delta_{21}\text{AGE} \\ &+ \delta_{22}\text{SMK} + \delta_{23}\text{RACE} + \delta_{24}\text{SEX}).\end{aligned}$$
10. Using the coding of the answer to Question 7, the revised odds ratio expression that compares urban with rural persons is $\exp(\beta_1 + \delta_{11}\text{AGE} + \delta_{12}\text{SMK} + \delta_{13}\text{RACE} + \delta_{14}\text{SEX})$.
11. The correct answer is b.

4 Maximum Likelihood Techniques: An Overview

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_4, © Springer Science+Business Media, LLC 2010

Introduction

In this chapter, we describe the general maximum likelihood (ML) procedure, including a discussion of likelihood functions and how they are maximized. We also distinguish between two alternative ML methods, the unconditional and the conditional approaches, and we give guidelines regarding how the applied user can choose between these methods. Finally, we provide a brief overview of how to make statistical inferences using ML estimates.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. Together with the objectives, this outline offers the user an overview of the content of this module. A detailed outline for review purposes follows the presentation.

I. Overview (page 106)

II. Background about maximum likelihood procedure (pages 106–107)

III. Unconditional vs. conditional methods (pages 107–111)

IV. The likelihood function and its use in the ML procedure (pages 111–117)

V. Overview on statistical inferences for logistic regression (pages 117–121)

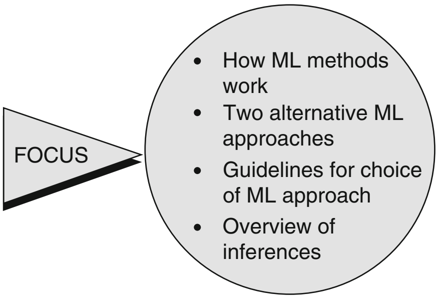
Objectives

Upon completing this chapter, the learner should be able to:

1. State or recognize when to use unconditional vs. conditional ML methods.
2. State or recognize what is a likelihood function.
3. State or recognize that the likelihood functions for unconditional vs. conditional ML methods are different.
4. State or recognize that unconditional vs. conditional ML methods require different computer programs.
5. State or recognize how an ML procedure works to obtain ML estimates of unknown parameters in a logistic model.
6. Given a logistic model, state or describe two alternative procedures for testing hypotheses about parameters in the model. In particular, describe each procedure in terms of the information used (log likelihood statistic or Z statistic) and the distribution of the test statistic under the null hypothesis (chi square or Z).
7. State, recognize, or describe three types of information required for carrying out statistical inferences involving the logistic model: the value of the maximized likelihood, the variance–covariance matrix, and a listing of the estimated coefficients and their standard errors.
8. Given a logistic model, state or recognize how interval estimates are obtained for parameters of interest; in particular, state that interval estimates are large sample formulae that make use of variance and covariances in the variance–covariance matrix.
9. Given a printout of ML estimates for a logistic model, use the printout information to describe characteristics of the fitted model. In particular, given such a printout, compute an estimated odds ratio for an exposure–disease relationship of interest.

Presentation

I. Overview


FOCUS
How ML methods work
Two alternative ML approaches
Guidelines for choice of ML approach
Overview of inferences

This presentation gives an overview of maximum likelihood (ML) methods as used in logistic regression analysis. We focus on how ML methods work, we distinguish between two alternative ML approaches, and we give guidelines regarding which approach to choose. We also give a brief overview on making statistical inferences using ML techniques.

II. Background About Maximum Likelihood Procedure

Maximum likelihood (ML) estimation

Least squares (LS) estimation: used in classical linear regression

- ML = LS when normality is assumed

Maximum likelihood (*ML*) *estimation* is one of several alternative approaches that statisticians have developed for estimating the parameters in a mathematical model. Another well-known and popular approach is *least squares* (*LS*) *estimation* which is described in most introductory statistics courses as a method for estimating the parameters in a classical straight line or multiple linear regression model. ML estimation and least squares estimation are different approaches that happen to give the same results for classical linear regression analyses when the dependent variable is assumed to be normally distributed.

ML estimation:

- Computer programs available
- General applicability
- Used for nonlinear models, e.g., the logistic model

For many years, ML estimation was not widely used because no computer software programs were available to carry out the complex calculations required. However, ML programs have been widely available in recent years. Moreover, when compared with least squares, the ML method can be applied in the estimation of complex nonlinear as well as linear models. In particular, because the logistic model is a nonlinear model, ML estimation is the preferred estimation method for logistic regression.

Discriminant function analysis:

- Previously used for logistic model
- Restrictive normality assumptions
- Gives biased results – odds ratio too high

Until the availability of computer software for ML estimation, the method used to estimate the parameters of a logistic model was *discriminant function analysis*. This method has been shown by statisticians to be essentially a least squares approach. Restrictive normality assumptions on the independent variables in the model are required to make statistical inferences about the model parameters. In particular, if any of the independent variables are dichotomous or categorical in nature, then the discriminant function method tends to give biased results, usually giving estimated odds ratios that are too high.

ML estimation:

- No restrictions on independent variables
- Preferred to discriminant analysis

ML estimation, on the other hand, requires no restrictions of any kind on the characteristics of the independent variables. Thus, when using ML estimation, the independent variables can be nominal, ordinal, and/or interval. Consequently, ML estimation is to be preferred over discriminant function analysis for fitting the logistic model.

III. Unconditional vs. Conditional Methods

Two alternative ML approaches:

1. Unconditional method
2. Conditional method

- Require different computer algorithms
- User must choose appropriate algorithm

There are actually two alternative ML approaches that can be used to estimate the parameters in a logistic model. These are called the *unconditional method* and the *conditional method*. These two methods require different computer algorithms. Thus, researchers using logistic regression modeling must decide which of these two algorithms is appropriate for their data. (See Computer Appendix.)

Computer Programs

SAS
SPSS
Stata

Three of the most widely available computer packages for unconditional ML estimation of the logistic model are SAS, SPSS, and Stata. Programs for conditional ML estimation are available in all three packages, but some are restricted to special cases. (See Computer Appendix.)

The Choice:

Unconditional – preferred if the number of parameters is *small* relative to the number of subjects

Conditional – preferred if the number of parameters is *large* relative to the number of subjects

In making the choice between *unconditional* and *conditional ML approaches*, the researcher needs to consider the number of parameters in the model relative to the total number of subjects under study. In general, *unconditional* ML estimation is preferred if the number of parameters in the model is *small* relative to the number of subjects. In contrast, *conditional* ML estimation is preferred if the number of parameters in the model is *large* relative to the number of subjects.

Small vs. large? debatable

Guidelines provided here

Exactly what is small vs. what is large is debatable and has not yet nor may ever be precisely determined by statisticians. Nevertheless, we can provide some guidelines for choosing the estimation method.

EXAMPLE: Unconditional Preferred

Cohort study: 10 year follow-up
$n = 700$
$D =$ CHD outcome
$E =$ exposure variable

$C_1, C_2, C_3, C_4, C_5 =$ covariables

$E \times C_1, E \times C_2, E \times C_3, E \times C_4, E \times C_5$
$=$ interaction terms

Number of parameters $= \enclose{circle}{12}$
(including intercept)

small relative to $n = \enclose{circle}{700}$

An example of a situation suitable for an unconditional ML program is a large cohort study that does not involve matching, for instance, a study of 700 subjects who are followed for 10 years to determine coronary heart disease status, denoted here as CHD. Suppose, for the analysis of data from such a study, a logistic model is considered involving an exposure variable E, five covariables C_1 through C_5 treated as confounders in the model, and five interaction terms of the form $E \times C_i$, where C_i is the ith covariable.

This model contains a total of 12 parameters, one for each of the variables plus one for the intercept term. Because the number of parameters here is 12 and the number of subjects is 700, this is a situation suitable for using *unconditional ML estimation*; that is, the number of parameters is *small* relative to the number of subjects.

EXAMPLE: Conditional Preferred

Case-control study
100 matched pairs
D = lung cancer

Matching variables:
age, race, sex, location

Other variables:
SMK (a confounder)
E (dietary characteristic)

In contrast, consider a case-control study involving 100 matched pairs. Suppose that the outcome variable is lung cancer and that controls are matched to cases on age, race, sex, and location. Suppose also that smoking status, a potential confounder denoted as SMK, is not matched but is nevertheless determined for both cases and controls, and that the primary exposure variable of interest, labeled as E, is some dietary characteristic, such as whether or not a subject has a high-fiber diet.

Logistic model for matching:

- uses dummy variables for matching strata
- 99 dummy variables for 100 strata
- E, SMK, and E × SMK also in model

Because the study design involves matching, a logistic model to analyze this data must control for the matching by using dummy variables to reflect the different matching strata, each of which involves a different matched pair. Assuming the model has an intercept, the model will need 99 dummy variables to incorporate the 100 matched pairs. Besides these variables, the model contains the exposure variable E, the covariable SMK, and perhaps even an interaction term of the form E × SMK.

Number of parameters =

$$\underset{\substack{\uparrow\\ \text{intercept}}}{1} + \underset{\substack{\uparrow\\ \text{dummy}\\ \text{variables}}}{99} + \underset{\substack{\uparrow\\ E,\ \text{SMK},\ E\times\text{SMK}}}{3} = \boxed{103}$$

large relative to 100 matched pairs ⇒ $\boxed{n = 200}$

To obtain the number of parameters in the model, we must count the one intercept, the coefficients of the 99 dummy variables, the coefficient of E, the coefficient of SMK, and the coefficient of the product term E × SMK. The total number of parameters is 103. Because there are 100 matched pairs in the study, the total number of subjects is, therefore, 200. This situation requires *conditional ML estimation* because the number of parameters, 103, is quite *large* relative to the number of subjects, 200.

REFERENCE

Chapter 11: Analysis of Matched Data Using Logistic Regression

A detailed discussion of logistic regression for matched data is provided in Chap. 11.

Guidelines:

The above examples indicate the following guidelines regarding the choice between unconditional and conditional ML methods or programs:

- Use *conditional* if matching

- Use *conditional ML estimation* whenever matching has been done; this is because the model will invariably be large due to the number of dummy variables required to reflect the matching strata.

- Use *unconditional* if no matching and number of variables not too large

- Use *unconditional ML estimation* if matching has not been done, provided the total number of variables in the model is not unduly large relative to the number of subjects.

EXAMPLE

Unconditional questionable if
- 10–15 confounders
- 10–15 product terms

Loosely speaking, this means that if the total number of confounders and the total number of interaction terms in the model are large, say 10–15 confounders and 10–15 product terms, the number of parameters may be getting too large for the unconditional approach to give accurate answers.

Safe rule:
Use *conditional* when in doubt.
- Gives unbiased results always.
- Unconditional may be biased (may overestimate odds ratios).

A safe rule is to use conditional ML estimation whenever in doubt about which method to use, because, theoretically, the conditional approach has been shown by statisticians to give unbiased results always. In contrast, the unconditional approach, when unsuitable, can give biased results and, in particular, can overestimate odds ratios of interest.

EXAMPLE: Conditional Required

Pair-matched case control study
measure of effect; OR

As a simple example of the need to use conditional ML estimation for matched data, consider again a pair-matched case-control study such as described above. For such a study design, the measure of effect of interest is an odds ratio for the exposure-disease relationship that adjusts for the variables being controlled.

EXAMPLE: (continued)

Assume only variables controlled are matched

Then

$$\underset{\text{biased}}{\underset{\uparrow}{\widehat{\text{OR}}_U}} = \underset{\text{correct}}{\underset{\uparrow}{(\widehat{\text{OR}}_C)}}^2$$

e.g.,

$$\widehat{\text{OR}}_C = 3 \Rightarrow \widehat{\text{OR}}_U = (3)^2 = 9$$

If the *only* variables being controlled are those involved in the matching, then the estimate of the odds ratio obtained by using unconditional ML estimation, which we denote by $\widehat{\text{OR}}_U$, is the square of the estimate obtained by using conditional ML estimation, which we denote by $\widehat{\text{OR}}_C$. Statisticians have shown that the correct estimate of this OR is given by the conditional method, whereas a biased estimate is given by the unconditional method.

Thus, for example, if the conditional ML estimate yields an estimated odds ratio of 3, then the unconditional ML method will yield a very large overestimate of 3 squared, or 9.

R-to-1 matching

$$\Downarrow$$

unconditional is overestimate of (correct) conditional estimate

More generally, whenever matching is used, even *R*-to-1 matching, where *R* is greater than 1, the unconditional estimate of the odds ratio that adjusts for covariables will give an overestimate, though not necessarily the square, of the conditional estimate.

Having now distinguished between the two alternative ML procedures, we are ready to describe the ML procedure in more detail and to give a brief overview of how statistical inferences are made using ML techniques.

IV. The Likelihood Function and Its Use in the ML Procedure

$L = L(\boldsymbol{\theta}) =$ likelihood function

$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)$

To describe the ML procedure, we introduce the likelihood function, *L*. This is a function of the unknown parameters in one's model and, thus, can alternatively be denoted as $L(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the collection of unknown parameters being estimated in the model. In matrix terminology, the collection $\boldsymbol{\theta}$ is referred to as a *vector*; its components are the individual parameters being estimated in the model, denoted here as θ_1, θ_2, up through θ_q, where *q* is the number of individual components.

E, V, W model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \sum_{i=1}^{p_1} \gamma_i V_i + E \sum_{j=1}^{p_2} \delta_j W_j$$

$$\boldsymbol{\theta} = (\alpha, \beta, \gamma_1, \gamma_2, \ldots, \delta_1, \delta_2, \ldots)$$

For example, using the *E, V, W* logistic model previously described and shown here again, the unknown parameters are α, β, the γ_is, and the δ_js. Thus, the vector of parameters $\boldsymbol{\theta}$ has α, β, the γ_is, and the δ_js as its components.

$L = L(\boldsymbol{\theta})$
= joint probability of observing the data

The *likelihood function L* or $L(\boldsymbol{\theta})$ *represents the joint probability or likelihood of observing the data that have been collected*. The term "joint probability" means a probability that combines the contributions of all the subjects in the study.

EXAMPLE

$n = 100$ trials
$p =$ probability of success
$x = 75$ successes
$n - x = 25$ failures

Pr (75 successes out of 100 trials) has binomial distribution

As a simple example, in a study involving 100 trials of a new drug, suppose the parameter of interest is the probability of a successful trial, which is denoted by p. Suppose also that, out of the n equal to 100 trials studied, there are x equal to 75 successful trials and $n - x$ equal to 25 failures. The probability of observing 75 successes out of 100 trials is a joint probability and can be described by the binomial distribution. That is, the model is a binomial-based model, which is different from and much less complex than the logistic model.

$$\Pr(X = 75 \mid \underset{\substack{\uparrow \\ \text{given}}}{} n = 100, p)$$

The binomial probability expression is shown here. This is stated as the probability that X, the number of successes, equals 75 given that there are n equal to 100 trials and that the probability of success on a single trial is p. Note that the vertical line within the probability expression means "given".

$$\Pr(X = 75 \mid n = 100, p)$$
$$= c \times p^{75} \times (1 - p)^{100 - 75}$$
$$= L(p)$$

This probability is numerically equal to a constant c times p to the 75th power times $1 - p$ to the $100 - 75$ or 25th power. This expression is the likelihood function for this example. It gives the probability of observing the results of the study as a function of the unknown parameters, in this case the single parameter p.

ML method maximizes the likelihood function $L(\boldsymbol{\theta})$

$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_q) =$ ML estimator

Once the likelihood function has been determined for a given set of study data, *the method of maximum likelihood chooses that estimator of the set of unknown parameters* $\boldsymbol{\theta}$ *which maximizes the likelihood function* $L(\boldsymbol{\theta})$. The estimator is denoted as $\hat{\boldsymbol{\theta}}$ and its components are $\hat{\theta}_1, \hat{\theta}_2$, and so on up through $\hat{\theta}_q$.

EXAMPLE (Binomial)

ML solution.
$\hat{p}$ maximizes
$L(p) = c \times p^{75} \times (1 - p)^{25}$

In the binomial example described above, the maximum likelihood solution gives that value of the parameter p which maximizes the likelihood expression c times p to the 75th power times $1 - p$ to the 25th power. The estimated parameter here is denoted as $\hat{p}$.

EXAMPLE (continued)

Maximum value obtained by solving

$$\frac{dL}{dp} = 0$$

for p:

$\hat{p} = 0.75$ "most likely"

$$p > \hat{p} = 0.75 \Rightarrow L(p) < \overset{\text{maximum}}{\overset{\downarrow}{L(p = 0.75)}}$$

e.g.,

$$p = 1 \Rightarrow L(1) = \overset{\text{binomial formula}}{c \times 1^{75} \times (1 - 1)^{25}}$$
$$= 0 < L(0.75)$$

$\hat{p} = 0.75 = \frac{75}{100}$, a sample proportion

Binomial model

$\Rightarrow \hat{p} = \frac{X}{n}$ is ML estimator

More complicated models $\Rightarrow$ complex calculations

The standard approach for maximizing an expression like the likelihood function for the binomial example here is to use calculus by setting the derivative dL/dp equal to 0 and solving for the unknown parameter or parameters.

For the binomial example, when the derivative dL/dp is set equal to 0, the ML solution obtained is $\hat{p}$ equal to 0.75. Thus, the value 0.75 is the "most likely" value for p in the sense that it maximizes the likelihood function L.

If we substitute into the expression for L a value for p exceeding 0.75, this will yield a smaller value for L than obtained when substituting p equal to 0.75. This is why 0.75 is called the ML estimator. For example, when p equals 1, the value for L using the binomial formula is 0, which is as small as L can get and is, therefore, less than the value of L when p equals the ML value of 0.75.

Note that for the binomial example, the ML value $\hat{p}$ equal to 0.75 is simply the sample proportion of the 100 trials that are successful. In other words, for a binomial model, *the sample proportion* always turns out to be the ML estimator of the parameter p. So for this model, it is not necessary to work through the calculus to derive this estimate. However, for models more complicated than the binomial, for example, the logistic model, calculus computations involving derivatives are required and are quite complex.

Maximizing $L(\boldsymbol{\theta})$ is equivalent to maximizing ln $L(\boldsymbol{\theta})$

Solve: $\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \theta_j} = 0, j = 1, 2, \ldots, q$

In general, maximizing the likelihood function $L(\boldsymbol{\theta})$ is equivalent to maximizing the natural log of $L(\boldsymbol{\theta})$, which is computationally easier. The components of $\boldsymbol{\theta}$ are then found as solutions of equations of partial derivatives as shown here. Each equation is stated as the partial derivative of the log of the likelihood function with respect to θ_j equals 0, where θ_j is the jth individual parameter.

q equations in q unknowns require *iterative* solution by computer

If there are q parameters in total, then the above set of equations is a set of q equations in q unknowns. These equations must then be solved iteratively, which is no problem with the right computer program.

Two alternatives:

Unconditional algorithm (L_U)
vs.
Conditional algorithm (L_C)
likelihoods

As described earlier, if the model is logistic, there are *two alternative types* of computer algorithms to choose from, an *unconditional* vs. a *conditional* algorithm. These algorithms use different likelihood functions, namely, L_U for the unconditional method and L_C for the conditional method.

Formula for L is built into computer algorithms

User inputs data and computer does calculations

The formulae for the likelihood functions for both the unconditional and conditional ML approaches are quite complex mathematically. The applied user of logistic regression, however, never has to see the formulae for L in practice because they are built into their respective computer algorithms. All the user has to do is learn how to input the data and to state the form of the logistic model being fit. Then the computer does the heavy calculations of forming the likelihood function internally and maximizing this function to obtain the ML solutions.

L formulae are different for unconditional and conditional methods

Although we do not want to emphasize the particular likelihood formulae for the unconditional vs. conditional methods, we do want to describe how these formulae are different. Thus, we briefly show these formulae for this purpose.

The unconditional formula: (a joint probability)

$$L_U = \overset{\text{cases}}{\overset{\downarrow}{\prod_{l=1}^{m_1} P(\mathbf{X}_l)}} \overset{\text{noncases}}{\overset{\downarrow}{\prod_{l=m_1+1}^{n} [1 - P(\mathbf{X}_l)]}}$$

$$P(\mathbf{X}) = \text{logistic model}$$
$$= \frac{1}{1 + e^{-(\alpha + \Sigma\beta_i X_i)}}$$

The *unconditional formula* is given first and directly describes the joint probability of the study data as the *product of the joint probability for the cases* (diseased persons) *and the joint probability for the noncases* (nondiseased persons). These two products are indicated by the large Π signs in the formula. We can use these products here by assuming that we have independent observations on all subjects. The probability of obtaining the data for the lth case is given by $P(\mathbf{X}_l)$, where $P(\mathbf{X})$ is the logistic model formula for individual $\mathbf{X}$. The probability of the data for the lth noncase is given by $1 - P(\mathbf{X}_l)$.

$$L_U = \frac{\prod_{l=1}^{n} \exp\left(\alpha + \sum_{i=1}^{k} \beta_i X_{il}\right)}{\prod_{l=1}^{n} \left[1 + \exp\left(\alpha + \sum_{i=1}^{k} \beta_i X_{il}\right)\right]}$$

When the logistic model formula involving the parameters is substituted into the likelihood expression above, the formula shown here is obtained after a certain amount of algebra is done. Note that this expression for the likelihood function L is a function of the unknown parameters α and the β_i.

The conditional formula:

$$L_C = \frac{\Pr(\text{observed data})}{\Pr(\text{all possible configurations})}$$

The conditional likelihood formula (L_C) reflects the probability of the observed data configuration relative to the probability of all possible configurations of the given data. To understand this, we describe the observed data configuration as a collection of m_1 cases and $n - m_1$ noncases. We denote the cases by the **X** vectors $\mathbf{X}_1$, $\mathbf{X}_2$, and so on through $\mathbf{X}_{m_1}$ and the noncases by $\mathbf{X}_{m_1+1}$, $\mathbf{X}_{m_1+2}$, through $\mathbf{X}_n$.

m_1 cases: $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{m_1})$
$n - m_1$ noncases:
$(\mathbf{X}_{m_1+1}, \mathbf{X}_{m_1+2}, \ldots, \mathbf{X}_n)$

L_C = Pr(first m_1 **X**s are cases | all possible configurations of **X**s)

The above configuration assumes that we have rearranged the observed data so that the m_1 cases are listed first and are then followed in listing by the $n - m_1$ noncases. Using this configuration, the conditional likelihood function gives the probability that the first m_1 of the observations actually go with the cases, given all possible configurations of the above n observations into a set of m_1 cases and a set of $n - m_1$ noncases.

EXAMPLE: Configurations

(1) Last m_1 **X**s are cases
$(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$
—— cases

(2) Cases of **X**s are in middle of listing
$(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$
—— cases

The term *configuration* here refers to one of the possible ways that the observed set of **X** vectors can be partitioned into m_1 cases and $n - m_1$ noncases. In example 1 here, for instance, the last m_1 **X** vectors are the cases and the remaining **X**s are noncases. In example 2, however, the m_1 cases are in the middle of the listing of all **X** vectors.

Possible configurations
= combinations of n things taken m_1 at a time
= $C^n_{m_1}$

The number of possible configurations is given by the number of combinations of n things taken m_1 at a time, which is denoted mathematically by the expression shown here, where the C in the expression denotes combinations.

$$L_C = \frac{\prod_{l=1}^{m_1} \mathrm{P}(\mathbf{X}_l) \prod_{l=m_1+1}^{n} [1 - \mathrm{P}(\mathbf{X}_l)]}{\sum_u \left\{ \prod_{l=1}^{m_1} \mathrm{P}(\mathbf{X}_{ul}) \prod_{l=m_1+1}^{n} [1 - \mathrm{P}(\mathbf{X}_{ul})] \right\}}$$

vs.

$$L_U = \prod_{l=1}^{m_1} \mathrm{P}(\mathbf{X}_l) \prod_{l=m_1+1}^{n} [1 - \mathrm{P}(\mathbf{X}_l)]$$

The formula for the conditional likelihood is then given by the expression shown here. The numerator is exactly the same as the likelihood for the unconditional method. The denominator is what makes the conditional likelihood different from the unconditional likelihood. Basically, the denominator sums the joint probabilities for all possible configurations of the m observations into m_1 cases and $n - m_1$ noncases. Each configuration is indicated by the u in the L_C formula.

$$L_{\mathrm{C}} = \frac{\prod_{l=1}^{m_1} \exp\left(\sum_{i=1}^{k} \beta_i X_{li}\right)}{\sum_{u} \left[\prod_{l=1}^{m_1} \exp\left(\sum_{i=1}^{k} \beta_i X_{lui}\right)\right]}$$

Note: α drops out of L_{C}

When the logistic model formula involving the parameters is substituted into the conditional likelihood expression above, the resulting formula shown here is obtained. This formula is not the same as the unconditional formula shown earlier. Moreover, in the conditional formula, the intercept parameter α has dropped out of the likelihood.

Conditional algorithm:

- Estimates βs
- Does not estimate α (nuisance parameter)

Note: OR involves only βs

The removal of the intercept α from the conditional likelihood is important because it means that when a conditional ML algorithm is used, estimates are obtained only for the β_i coefficients in the model and not for α. Because the usual focus of a logistic regression analysis is to estimate an odds ratio, which involves the βs and not α, we usually do not care about estimating α and, therefore, consider α to be a nuisance parameter.

Case-control study:
cannot estimate α

In particular, if the data come from a case-control study, we cannot estimate α because we cannot estimate risk, and the conditional likelihood function does not allow us to obtain any such estimate.

$L_{\mathrm{U}} \neq L_{\mathrm{C}}$

direct joint probability

does not require estimating nuisance parameters

Regarding likelihood functions, then, we have shown that the unconditional and conditional likelihood functions involve different formulae. The unconditional formula has the theoretical advantage in that it is developed directly as a joint probability of the observed data. The conditional formula has the advantage that it does not require estimating nuisance parameters like α.

Stratified data, e.g., matching,
⇓
many nuisance parameters

100 nuisance parameters

are not estimated using L_{C}

unnecessarily estimated using L_{U}

If the data are stratified, as, for example, by matching, it can be shown that there are as many nuisance parameters as there are matched strata. Thus, for example, if there are 100 matched pairs, then 100 nuisance parameters do not have to be estimated when using conditional estimation, whereas these 100 parameters would be unnecessarily estimated when using unconditional estimation.

Matching:

Unconditional ⇒ biased estimates of βs
Conditional ⇒ unbiased estimates of βs

If we consider the other parameters in the model for matched data, that is, the βs, the unconditional likelihood approach gives biased estimates of the βs, whereas the conditional approach gives unbiased estimates of the βs.

V. Overview on Statistical Inferences for Logistic Regression

Chap. 5: Statistical Inferences Using Maximum Likelihood Techniques

We have completed our description of the ML method in general, distinguished between unconditional and conditional approaches, and distinguished between their corresponding likelihood functions. We now provide a brief overview of how statistical inferences are carried out for the logistic model. A detailed discussion of statistical inferences is given in the next chapter.

Statistical inferences involve the following:

- Testing hypotheses
- Obtaining confidence intervals

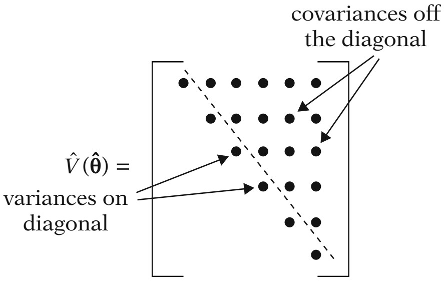
Once the ML estimates have been obtained, the next step is to use these estimates to make *statistical inferences* concerning the exposure–disease relationships under study. This step includes testing hypotheses and obtaining confidence intervals for parameters in the model.

Quantities required from computer output:

Inference-making can be accomplished through the use of two quantities that are part of the output provided by standard ML estimation programs.

1. Maximized likelihood value $L(\hat{\boldsymbol{\theta}})$
2. Estimated variance–covariance matrix

The first of these quantities is the *maximized likelihood value*, which is simply the numerical value of the likelihood function L when the ML estimates ($\hat{\boldsymbol{\theta}}$) are substituted for their corresponding parameter values (θ). This value is called $L(\hat{\boldsymbol{\theta}})$ in our earlier notation.


covariances off the diagonal
$\hat{V}(\hat{\boldsymbol{\theta}})$ =
variances on diagonal

Note: $\widehat{\text{cov}}(\hat{\theta}_1, \hat{\theta}_2) = r_{12}s_1s_2$

The second quantity is the *estimated variance–covariance matrix*. This matrix, $\hat{V}$ of $\hat{\boldsymbol{\theta}}$, has as its diagonal the estimated variances of each of the ML estimates. The values off the diagonal are the covariances of pairs of ML estimates. The reader may recall that the covariance between two estimates is the correlation times the standard error of each estimate.

Importance of $\hat{V}(\hat{\boldsymbol{\theta}})$:

inferences require accounting for variability and covariability

The variance–covariance matrix is important because the information contained in it is used in the computations required for hypothesis testing and confidence interval estimation.

(3) Variable listing

Variable	ML Coefficient	S. E.
Intercept	$\hat{\alpha}$	$s_{\hat{\alpha}}$
X_1	$\hat{\beta}_1$	$s_{\hat{\beta}_1}$
•	•	•
•	•	•
•	•	•
X_k	$\hat{\beta}_k$	$s_{\hat{\beta}_k}$

In addition to the maximized likelihood value and the variance–covariance matrix, other information is also provided as part of the output. This information typically includes, as shown here, *a listing of each variable followed by its ML estimate and standard error*. This information provides another way to carry out hypothesis testing and interval estimation. Moreover, this listing gives the primary information used for calculating odds ratio estimates and predicted risks. The latter can only be done, however, if the study has a follow-up design.

EXAMPLE

Cohort study – Evans Country, GA

$n = 609$ white males
9-year follow-up
$D =$ CHD status

Output: $-2 \ln \hat{L} = 347.23$

	Variable	ML Coefficient	S. E.
	Intercept	−4.0497	−1.2550
	CAT	−12.6894	3.1047
	AGE	0.0350	0.0161
	CHL	0.0055	0.0042
V_s	ECG	0.3671	0.3278
	SMK	0.7732	0.3273
	HPT	1.0466	0.3316
	CC	0.0692	0.3316
	CH	2.3318	0.7427

CC = CAT × CHL and CH = CAT × HPT

Ws

An example of ML computer output giving the above information is provided here. This output considers study data on a cohort of 609 white males in Evans County, Georgia, who were followed for 9 years to determine coronary heart disease (CHD) status. The output considers a logistic model involving eight variables, which are denoted as CAT (catecholamine level), AGE, CHL (cholesterol level), ECG (electrocardiogram abnormality status), SMK (smoking status), HPT (hypertension status), CC, and CH. The latter two variables are product terms of the form CC = CAT × CHL and CH = CAT × HPT.

The exposure variable of interest here is the variable CAT, and the five covariables of interest, that is, the Cs are AGE, CHL, ECG, SMK, and HPT. Using our E, V, W model framework introduced in Chapter 2, we have E equals CAT, the five covariables equal to the Vs, and two W variables, namely, CHL and HPT.

The output information includes −2 times the natural log of the maximized likelihood value, which is 347.23, and a listing of each variable followed by its ML estimate and standard error. We will show the variance–covariance matrix shortly.

EXAMPLE (continued)

$\widehat{\text{OR}}$ considers coefficients of CAT, CC, and CH

$\widehat{\text{OR}} = \exp(\hat{\beta} + \hat{\delta}_1\text{CHL} + \hat{\delta}_2\text{HPT})$

where

$\hat{\beta} = -12.6894$

$\hat{\delta}_1 = 0.0692$

$\hat{\delta}_2 = -2.3318$

$$\widehat{\text{OR}} = \exp[-12.6894 + 0.0692\ \text{CHL} + (-2.3318)\text{HPT}]$$

Must specify:

CHL and HPT

effect modifiers

Note. $\widehat{\text{OR}}$ different for different values specified for CHL and HPT

		HPT	
		0	1
CHL	200	3.16	0.31
	220	12.61	1.22
	240	50.33	4.89

CHL = 200, HPT = 0: $\widehat{\text{OR}} = 3.16$

CHL = 220, HPT = 1: $\widehat{\text{OR}} = 1.22$

$\widehat{\text{OR}}$ adjusts for AGE, CHL, ECG, SMK, and HPT (the *V* variables)

We now consider how to use the information provided to obtain an estimated odds ratio for the fitted model. Because this model contains the product terms CC equal to CAT × CHL, and CH equal to CAT × HPT, the estimated odds ratio for the effect of CAT must consider the coefficients of these terms as well as the coefficient of CAT.

The formula for this estimated odds ratio is given by the exponential of the quantity $\hat{\beta}$ plus $\hat{\delta}_1$ times CHL plus $\hat{\delta}_2$ times HPT, where $\hat{\beta}$ equals -12.6894 is the coefficient of CAT, $\hat{\delta}_1$ equals 0.0692 is the coefficient of the interaction term CC, and $\hat{\delta}_2$ equals -2.3318 is the coefficient of the interaction term CH.

Plugging the estimated coefficients into the odds ratio formula yields the expression: *e* to the quantity -12.6894 plus 0.0692 times CHL plus -2.3318 times HPT.

To obtain a numerical value from this expression, it is necessary to specify a value for CHL and a value for HPT. Different values for CHL and HPT will, therefore, yield different odds ratio values, as should be expected because the model contains interaction terms.

The table shown here illustrates different odds ratio estimates that can result from specifying different values of the effect modifiers. In this table, the values of CHL are 200, 220, and 240; the values of HPT are 0 and 1, where 1 denotes a person who has hypertension. The cells within the table give the estimated odds ratios computed from the above expression for the odds ratio for different combinations of CHL and HPT.

For example, when CHL equals 200 and HPT equals 0, the estimated odds ratio is given by 3.16; when CHL equals 220 and HPT equals 1, the estimated odds ratio is 1.22. Note that each of the estimated odds ratios in this table describes the association between CAT and CHD adjusted for the five covariables AGE, CHL, ECG, SMK, and HPT because each of the covariables is contained in the model as *V* variables.

$\widehat{\text{OR}}$s = point estimators

Variability of $\widehat{\text{OR}}$ considered for statistical inferences

The estimated model coefficients and the corresponding odds ratio estimates that we have just described are point estimates of unknown population parameters. Such point estimates have a certain amount of variability associated with them, as illustrated, for example, by the standard errors of each estimated coefficient provided in the output listing. We consider the variability of our estimates when we make statistical inferences about parameters of interest.

Two types of inferences:

(1) Testing hypotheses

(2) Interval estimation

We can use two kinds of inference-making procedures. One is testing hypotheses about certain parameters; the other is deriving interval estimates of certain parameters.

EXAMPLES

(1) Test for H_0: OR = 1

(2) Test for significant interaction, e.g., $\delta_1 \neq 0$?

(3) Interval estimate: 95% confidence interval for $\text{OR}_{\text{CAT, CHD}}$ controlling for 5 *V*s and 2 *W*s

Interaction: must specify *W*s
e.g., 95% confidence interval when CAT = 220 and HPT = 1

As an example of a test, we may wish to test the null hypothesis that an odds ratio is equal to the null value.

Or, as another example, we may wish to test for evidence of significant interaction, for instance, whether one or more of the coefficients of the product terms in the model are significantly nonzero.

As an example of an interval estimate, we may wish to obtain a 95% confidence interval for the adjusted odds ratio for the effect of CAT on CHD, controlling for the five *V* variables and the two *W* variables. Because this model contains interaction terms, we need to specify the values of the *W*s to obtain numerical values for the confidence limits. For instance, we may want the 95% confidence interval when CHL equals 220 and HPT equals 1.

Two testing procedures:

(1) *Likelihood ratio test*: a chi-square statistic using $-2 \ln \hat{L}$.

(2) *Wald test*: a *Z* test using standard errors listed with each variable.

Note: Since Z^2 is $\chi^2_{1\text{ df}}$, the Wald test can equivalently be considered a chi-square test.

When using ML estimation, we can carry out hypothesis testing by using one of two procedures, the *likelihood ratio test* and the *Wald test*. The likelihood ratio test is a chi-square test that makes use of maximized likelihood values such as those shown in the output. The Wald test is a *Z* test; that is, the test statistic is approximately standard normal. The Wald test makes use of the standard errors shown in the listing of variables and associated output information. Each of these procedures is described in detail in the next chapter.

Large samples: both procedures give approximately the same results

Small or moderate samples: different results possible; likelihood ratio test preferred

Both testing procedures should give approximately the *same answer in large samples but may give different results in small or moderate samples*. In the latter case, statisticians prefer the likelihood ratio test to the Wald test.

Confidence intervals

- use large sample formulae
- use variance–covariance matrix

Confidence intervals are carried out by using large sample formulae that make use of the information in the variance–covariance matrix, which includes the variances of estimated coefficients together with the covariances of pairs of estimated coefficients.

EXAMPLE $\hat{V}(\hat{\boldsymbol{\theta}})$

	Intercept	CAT	AGE	CC	CH
Intercept	1.5750	−0.6629	−0.0136	0.0034	0.0548
CAT		9.6389	−0.0021	−0.0437	−0.0049
AGE			0.0003	0.0000	−0.0010
•			•		•
•				•	•
•				•	•
CC				0.0002	−0.0016
CH					0.5516

An example of the estimated variance–covariance matrix is given here. Note, for example, that the variance of the coefficient of the CAT variable is 9.6389, the variance for the CC variable is 0.0002, and the covariance of the coefficients of CAT and CC is −0.0437.

No interaction: variance only

Interaction: variances and covariances

If the model being fit contains no interaction terms and if the exposure variable is a (0, 1) variable, then only a variance estimate is required for computing a confidence interval. If the model contains interaction terms, then both variance and covariance estimates are required; in this latter case, the computations required are much more complex than when there is no interaction.

SUMMARY

Chapters up to this point:

1. Introduction
2. Important Special Cases
3. Computing the Odds Ratio
4. ✓ ML Techniques: An Overview

This presentation is now complete. In summary, we have described how ML estimation works, have distinguished between unconditional and conditional methods and their corresponding likelihood functions, and have given an overview of how to make statistical inferences using ML estimates.

We suggest that the reader review the material covered here by reading the summary outline that follows. Then you may work the practice exercises and test.

5. Statistical Inferences Using ML Techniques

In the next chapter, we give a detailed description of how to carry out both testing hypotheses and confidence interval estimation for the logistic model.

Detailed Outline

I. **Overview** (page 106)

Focus:

- How ML methods work
- Two alternative ML approaches
- Guidelines for choice of ML approach
- Overview of statistical inferences

II. **Background about maximum likelihood procedure** (pages 106–107)

A. Alternative approaches to estimation: least squares (LS), maximum likelihood (ML), and discriminant function analysis.

B. ML is now the preferred method – computer programs now available; general applicability of ML method to many different types of models.

III. **Unconditional vs. conditional methods** (pages 107–111)

A. Require different computer programs; user must choose appropriate program.

B. *Unconditional* preferred if number of parameters *small* relative to number of subjects, whereas *conditional* preferred if number of parameters *large* relative to number of subjects.

C. Guidelines: use conditional if matching; use unconditional if no matching and number of variables not too large; when in doubt, use conditional – always unbiased.

IV. **The likelihood function and its use in the ML procedure** (pages 111–117)

A. $L = L(\boldsymbol{\theta})$ = likelihood function; gives joint probability of observing the data as a function of the set of unknown parameters given by $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)$.

B. ML method maximizes the likelihood function $L(\boldsymbol{\theta})$.

C. ML solutions solve a system of q equations in q unknowns; this system requires an *iterative* solution by computer.

D. Two alternative likelihood functions for logistic regression: unconditional (L_{U}) and conditional (L_{C}); formulae are built into unconditional and conditional computer algorithms.

E. User inputs data and computer does calculations.

F. Conditional likelihood reflects the probability of observed data configuration relative to the probability of all possible configurations of the data.

G. Conditional algorithm estimates βs but not α (nuisance parameter).

H. Matched data: unconditional gives biased estimates, whereas conditional gives unbiased estimates.

V. **Overview on statistical inferences for logistic regression** (pages 117–121)

A. Two types of inferences: testing hypotheses and confidence interval estimation.

B. Three items obtained from computer output for inferences:

i. Maximized likelihood value $L(\hat{\boldsymbol{\theta}})$

ii. Estimated variance–covariance matrix $\hat{V}(\hat{\boldsymbol{\theta}})$: variances on diagonal and covariances on the off-diagonal;

iii. Variable listing with ML estimates and standard errors.

C. Two testing procedures:

i. *Likelihood ratio test*: a chi-square statistic using $-2 \ln \hat{L}$.

ii. *Wald test*: a Z test (or equivalent χ^2 test) using standard errors listed with each variable.

D. Both testing procedures give approximately same results with large samples; with small samples, different results are possible; likelihood ratio test is preferred.

E. Confidence intervals: use large sample formulae that involve variances and covariances from variance–covariance matrix.

Practice Exercises

True or False (Circle T or F)

T F 1. When estimating the parameters of the logistic model, least squares estimation is the preferred method of estimation.

T F 2. Two alternative maximum likelihood approaches are called unconditional and conditional methods of estimation.

T F 3. The conditional approach is preferred if the number of parameters in one's model is small relative to the number of subjects in one's data set.

T F 4. Conditional ML estimation should be used to estimate logistic model parameters if matching has been carried out in one's study.

T F 5. Unconditional ML estimation gives unbiased results always.

T F 6. The likelihood function L(**θ**) represents the joint probability of observing the data that has been collected for analysis.

T F 7. The maximum likelihood method maximizes the function ln L(**θ**).

T F 8. The likelihood function formulae for both the unconditional and conditional approaches are the same.

T F 9. The maximized likelihood value $L(\hat{\boldsymbol{\theta}})$ is used for confidence interval estimation of parameters in the logistic model.

T F 10. The likelihood ratio test is the preferred method for testing hypotheses about parameters in the logistic model.

Test

True or False (Circle T or F)

T F 1. Maximum likelihood estimation is preferred to least squares estimation for estimating the parameters of the logistic and other nonlinear models.

T F 2. If discriminant function analysis is used to estimate logistic model parameters, biased estimates can be obtained that result in estimated odds ratios that are too high.

T F 3. In a case-control study involving 1,200 subjects, a logistic model involving 1 exposure variable, 3 potential confounders, and 3 potential effect modifiers is to be estimated. Assuming no matching has been done, the preferred method of estimation for this model is conditional ML estimation.

T F 4. Until recently, the most widely available computer packages for fitting the logistic model have used unconditional procedures.

T F 5. In a matched case-control study involving 50 cases and 2-to-1 matching, a logistic model used to analyze the data will contain a small number of parameters relative to the total number of subjects studied.

T F 6. If a likelihood function for a logistic model contains ten parameters, then the ML solution solves a system of ten equations in ten unknowns by using an iterative procedure.

T F 7. The conditional likelihood function reflects the probability of the observed data configuration relative to the probability of all possible configurations of the data.

T F 8. The nuisance parameter α is not estimated using an unconditional ML program.

T F 9. The likelihood ratio test is a chi-square test that uses the maximized likelihood value $\hat{L}$ in its computation.

T F 10. The Wald test and the likelihood ratio test of the same hypothesis give approximately the same results in large samples.

T F 11. The variance–covariance matrix printed out for a fitted logistic model gives the variances of each variable in the model and the covariances of each pair of variables in the model.

T F 12. Confidence intervals for odds ratio estimates obtained from the fit of a logistic model use large sample formulae that involve variances and possibly covariances from the variance–covariance matrix.

The printout given below comes from a matched case-control study of 313 women in Sydney, Australia (Brock et al., 1988), to assess the etiologic role of sexual behaviors and dietary factors on the development of cervical cancer. Matching was done on age and socioeconomic status. The outcome variable is cervical cancer status (yes/no), and the independent variables considered here (all coded as 1, 0) are vitamin C intake (VITC, high/low), the number of lifetimesexual partners (NSEX, high/low), age at first intercourse (SEXAGE, old/young), oral contraceptive pill use (PILLM ever/never), and smoking status (CSMOK, ever/never).

Variable	Coefficient	S.E.	e^{Coeff}	P	95% Conf. Int. for e^{Coeff}	
VITC	−0.24411	0.14254	0.7834	.086	0.5924	1.0359
NSEX	0.71902	0.16848	2.0524	.000	1.4752	2.8555
SEXAGE	−0.19914	0.25203	0.8194	.426	0.5017	1.3383
PILLM	0.39447	0.19004	1.4836	.037	1.0222	2.1532
CSMOK	1.59663	0.36180	4.9364	.000	2.4290	10.0318

MAX LOG LIKELIHOOD = −73.5088

Using the above printout, answer the following questions:

13. What method of estimation should have been used to fit the logistic model for this data set? Explain.
14. Why don't the variables age and socioeconomic status appear in the printout?
15. Describe how to compute the odds ratio for the effect of pill use in terms of an estimated regression coefficient in the model. Interpret the meaning of this odds ratio.
16. What odds ratio is described by the value *e* to −0.24411? Interpret this odds ratio.
17. State two alternative ways to describe the null hypothesis appropriate for testing whether the odds ratio described in Question 16 is significant.
18. What is the 95% confidence interval for the odds ratio described in Question 16, and what parameter is being estimated by this interval?
19. The *P*-values given in the table correspond to Wald test statistics for each variable adjusted for the others in the model. The appropriate *Z* statistic is computed by dividing the estimated coefficient by its standard error. What is the *Z* statistic corresponding to the *P*-value of .086 for the variable VITC?
20. For what purpose is the quantity denoted as MAX LOG LIKELIHOOD used?

Answers to Practice Exercises

1. F: ML estimation is preferred
2. T
3. F: conditional is preferred if number of parameters is large
4. T
5. F: conditional gives unbiased results
6. T
7. T
8. F: L_U and L_C are different
9. F: The variance–covariance matrix is used for confidence interval estimation
10. T

5 Statistical Inferences Using Maximum Likelihood Techniques

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_5, © Springer Science+Business Media, LLC 2010

Introduction

We begin our discussion of statistical inference by describing the computer information required for making inferences about the logistic model. We then introduce examples of three logistic models that we use to describe hypothesis testing and confidence interval estimation procedures. We consider models with no interaction terms first, and then we consider how to modify procedures when there is interaction. Two types of testing procedures are given, namely, the likelihood ratio test and the Wald test. Confidence interval formulae are provided that are based on large sample normality assumptions. A final review of all inference procedures is described by way of a numerical example.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

I. Overview (page 132)

II. Information for making statistical inferences (pages 132–133)

III. Models for inference-making (pages 133–134)

IV. The likelihood ratio test (pages 134–138)

V. The Wald test (pages 138–140)

VI. Interval estimation: one coefficient (pages 140–142)

VII. Interval estimation: interaction (pages 142–146)

VIII. Numerical example (pages 146–153)

Objectives

Upon completion of this chapter, the learner should be able to:

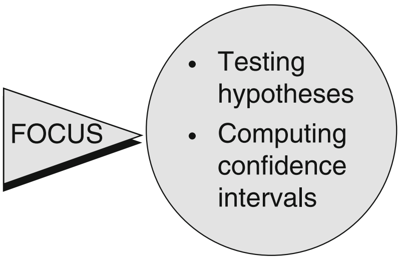
1. State the *null hypothesis* for testing the significance of a collection of one or more variables in terms of regression coefficients of a given logistic model.
2. Describe how to carry out a *likelihood ratio test* for the significance of one or more variables in a given logistic model.
3. Use computer information for a fitted logistic model to carry out a likelihood ratio test for the significance of one or more variables in the model.
4. Describe how to carry out a *Wald test* for the significance of a single variable in a given logistic model.
5. Use computer information for a fitted logistic model to carry out a Wald test for the significance of a single variable in the model.
6. Describe how to compute a *95% confidence interval* for an odds ratio parameter that can be estimated from a given logistic model when
   a. The model contains no interaction terms
   b. The model contains interaction terms
7. Use computer information for a fitted logistic model to compute a 95% confidence interval for an odds ratio expression estimated from the model when
   a. The model contains no interaction terms
   b. The model contains interaction terms

Presentation

I. Overview

Previous chapter:

- How ML methods work
- Unconditional vs. conditional approaches



In the previous chapter, we described how ML methods work in general and we distinguished between two alternative approaches to estimation – the unconditional and the conditional approach.

In this chapter, we describe how statistical inferences are made using ML techniques in logistic regression analyses. We focus on procedures for testing hypotheses and computing confidence intervals about logistic model parameters and odds ratios derived from such parameters.

II. Information for Making Statistical Inferences

Quantities required from output:

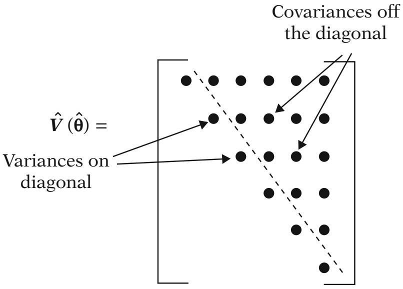
Once ML estimates have been obtained, these estimates can be used to make statistical inferences concerning the exposure–disease relationships under study. Three quantities are required from the output provided by standard ML estimation programs.

(1) Maximized likelihood value: $L(\hat{\boldsymbol{\theta}})$

The first of these quantities is the *maximized likelihood value*, which is the numerical value of the likelihood function L when the ML estimates are substituted for their corresponding parameter values; this value is called L of $\hat{\boldsymbol{\theta}}$ in our earlier notation.

(2) Estimated variance–covariance matrix: $\hat{V}(\hat{\boldsymbol{\theta}})$

The second quantity is the *estimated variance–covariance matrix*, which we denote as $\hat{V}$ of $\hat{\boldsymbol{\theta}}$.


Covariances off the diagonal
$\hat{V}(\hat{\boldsymbol{\theta}})$ =
Variances on diagonal

The estimated variance–covariance matrix has on its diagonal the estimated variances of each of the ML estimates. The values off the diagonal are the covariances of paris of ML estimates.

$\widehat{\text{cov}}(\hat{\theta}_1, \hat{\theta}_2) = r_{12}s_1s_2$

The reader may recall that the covariance between two estimates is the correlation times the standard errors of each estimate.

Importance of $\hat{V}(\hat{\boldsymbol{\theta}})$:

Inferences require variances and covariances

The variance–covariance matrix is important because hypothesis testing and confidence interval estimation require variances and sometimes covariances for computation.

(3) Variable listing:

Variable	ML Coefficient	S.E.
Intercept	$\hat{\alpha}$	$s_{\hat{\alpha}}$
X_1	$\hat{\beta}_1$	$s_{\hat{\beta}_1}$
•	•	•
•	•	•
•	•	•
X_k	$\hat{\beta}_k$	$s_{\hat{\beta}_K}$

In addition to the maximized likelihood value and the variance–covariance matrix, other information is also provided as part of the output. This typically includes, as shown here, *a listing of each variable followed by its ML estimate and standard error*. This information provides another way of carrying out hypothesis testing and confidence interval estimation, as we will describe shortly. Moreover, this listing gives the primary information used for calculating odds ratio estimates and predicted risks. The latter can only be done, however, provided the study has a follow-up type of design.

III. Models for Inference-Making

Model 1: $\text{logit}\, P_1(X) = \alpha + \beta_1X_1 + \beta_2X_2$

Model 2: $\text{logit}\, P_2(X) = \alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$

Model 3: $\text{logit}\, P_3(X) = \alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_1X_3 + \beta_5X_2X_3$

To illustrate how statistical inferences are made using the above information, we consider the following three models, each written in logit form. Model 1 involves two variables X_1 and X_2. Model 2 contains these same two variables and a third variable X_3. Model 3 contains the same three X's as in model 2 plus two additional variables, which are the product terms X_1X_3 and X_2X_3.

$\hat{L}_1$, $\hat{L}_2$, $\hat{L}_3$ are $\hat{L}$s for models 1–3

Let $\hat{L}_1$, $\hat{L}_2$, and $\hat{L}_3$ denote the maximized likelihood values based on fitting Models 1, 2, and 3, respectively. Note that the fitting may be done either by unconditional or conditional methods, depending on which method is more appropriate for the model and data set being considered.

$\hat{L}_1 \leq \hat{L}_2 \leq \hat{L}_3$

Because the more parameters a model has, the better it fits the data, it follows that $\hat{L}_1$ must be less than or equal to $\hat{L}_2$, which, in turn, must be less than or equal to $\hat{L}_3$.

$\hat{L}$ similar to R^2

This relationship among the $\hat{L}$s is similar to the property in classical multiple linear regression analyses that the more parameters a model has, the higher is the R-square statistic for the model. In other words, the maximized likelihood value $\hat{L}$ is similar to R-square, in that the higher the $\hat{L}$, the better the fit.

$\ln \hat{L}_1 \leq \ln \hat{L}_2 \leq \ln \hat{L}_3$

It follows from algebra that if $\hat{L}_1$ is less than or equal to $\hat{L}_2$, which is less than $\hat{L}_3$, then the same inequality relationship holds for the natural logarithms of these $\hat{L}$s.

$-2 \ln \hat{L}_3 \leq -2 \ln \hat{L}_2 \leq -2 \ln \hat{L}_1$

However, if we multiply each log of $\hat{L}$ by -2, then the inequalities switch around so that $-2 \ln \hat{L}_3$ is less than or equal to $-2 \ln \hat{L}_2$, which is less than $-2 \ln \hat{L}_1$.

$-2 \ln \hat{L} =$ log likelihood statistic used in likelihood ratio (LR) test

The statistic $-2 \ln \hat{L}_1$ is called the *log likelihood statistic* for Model 1, and similarly, the other two statistics are the log likelihood statistics for their respective models. These statistics are important because they can be used to test hypotheses about parameters in the model using what is called a *likelihood ratio test*, which we now describe.

IV. The Likelihood Ratio Test

$-2 \ln L_1 - (-2 \ln L_2) = \text{LR}$
is approximate chi square

df = difference in number of parameters (degrees of freedom)

Statisticians have shown that the difference between log likelihood statistics for two models, one of which is a special case of the other, has an approximate chisquare distribution in large samples. Such a test statistic is called a *likelihood ratio* or *LR* statistic. The degrees of freedom (df) for this chi-square test are equal to the difference between the number of parameters in the two models.

Model 1: $\text{logit } P_1(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2$
Model 2: $\text{logit } P_2(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
Note. special case = subset

Model 1 special case of Model 2
Model 2 special case of Model 3

Note that one model is considered a special case of another if one model contains a subset of the parameters in the other model. For example, Model 1 above is a special case of Model 2; also, Model 2 is a special case of Model 3.

LR statistic (like F statistic) compares two models:

Full model = larger model

Reduced model = smaller model

In general, the likelihood ratio statistic, like an F statistic in classical multiple linear regression, requires the identification of two models to be compared, one of which is a special case of the other. The larger model is sometimes called the *full model* and the smaller model is sometimes called the *reduced model*; that is, the reduced model is obtained by setting certain parameters in the full model equal to zero.

H_0: parameters in full model equal to zero

df = number of parameters set equal to zero

The set of parameters in the full model that is set equal to zero specify the null hypothesis being tested. Correspondingly, the degrees of freedom for the likelihood ratio test are equal to the number of parameters in the larger model that must be set equal to zero to obtain the smaller model.

EXAMPLE

Model 1 vs. Model 2

Model 2 (full model):

$$\text{logit } P_2(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Model 1 (reduced model):

$$\text{logit } P_1(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2$$

H_0: $\beta_3 = 0$ (similar to partial F)

As an example of a likelihood ratio test, let us now compare Model 1 with Model 2. Because Model 2 is the larger model, we can refer to Model 2 as the full model and to Model 1 as the reduced model. The additional parameter in the full model that is not part of the reduced model is β_3, the coefficient of the variable X_3. Thus, the null hypothesis that compares Models 1 and 2 is stated as β_3 equal to 0. This is similar to the null hypothesis for a partial F test in classical multiple linear regression analysis.

Model 2:

$$\text{logit } P_2(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Suppose $X_3 = E(0, 1)$ and X_1, X_2 are confounders.

Then $\text{OR} = e^{\beta_3}$

Now consider Model 2, and suppose that the variable X_3 is a (0, 1) exposure variable E and that the variables X_1 and X_2 are confounders. Then the odds ratio for the exposure–disease relationship that adjusts for the confounders is given by e to β_3.

H_0: $\beta_3 = 0 \Leftrightarrow H_0$: $\text{OR} = e^0 = 1$

Thus, in this case, testing the null hypothesis that β_3 equals 0 is equivalent to testing the null hypothesis that the adjusted odds ratio for the effect of exposure is equal to e to 0 or 1.

$\text{LR} = -2 \ln \hat{L}_1 - (-2 \ln \hat{L}_2)$

To test this null hypothesis, the corresponding likelihood ratio statistic is given by the difference $-2 \ln \hat{L}_1$ minus $-2 \ln \hat{L}_2$.

EXAMPLE (continued)

Ratio of likelihoods

$$-2\ln\hat{L}_1-(-2\ln\hat{L}_2)=-2\ln\left(\frac{\hat{L}_1}{\hat{L}_2}\right)$$

LR approximate χ^2 variable with df = 1 if n large

Algebraically, this difference can also be written as -2 times the natural log of the ratio of $\hat{L}_1$ divided by $\hat{L}_2$, shown on the right-hand side of the equation here. This latter version of the test statistic is a ratio of maximized likelihood values; this explains why the test is called the likelihood ratio test.

The likelihood ratio statistic for this example has approximately a chi-square distribution if the study size is large. The degrees of freedom for the test is one because, when comparing Models 1 and 2, only one parameter, namely, β_3, is being set equal to zero under the null hypothesis.

How the LR test works:

If X_3 makes a large contribution, then $\hat{L}_2$ much greater than $\hat{L}_1$

We now describe how the likelihood ratio test works and why the test statistic is approximately chi square. We consider what the value of the test statistic would be if the additional variable X_3 makes an extremely large contribution to the risk of disease over that already contributed by X_1 and X_2. Then, it follows that the maximized likelihood value $\hat{L}_2$ is much larger than the maximized likelihood value $\hat{L}_1$.

If $\hat{L}_2$ much larger than $\hat{L}_1$, then

$$\frac{\hat{L}_1}{\hat{L}_2}\approx 0$$

If $\hat{L}_2$ is much larger than $\hat{L}_1$, then the ratio $\hat{L}_1$ divided by $\hat{L}_2$ becomes a very small fraction; that is, this ratio approaches 0.

[*Note*. $\ln_e$ (fraction) = negative]

$$\Rightarrow \ln\left(\frac{\hat{L}_1}{\hat{L}_2}\right)\approx \ln(0)=-\infty$$

Now the natural log of any fraction between 0 and 1 is a negative number. As this fraction approaches 0, the log of the fraction, which is negative, approaches the log of 0, which is $-\infty$.

$$\Rightarrow \text{LR}=-2\ln\left(\frac{\hat{L}_1}{\hat{L}_2}\right)\approx\infty$$

Thus, X_3 highly significant $\Rightarrow$ LR large and positive.

If we multiply the log likelihood ratio by -2, we then get a number that approaches $+\infty$. Thus, the likelihood ratio statistic for a highly significant X_3 variable is large and positive and approaches $+\infty$. This is exactly the type of result expected for a chi-square statistic.

If X_3 makes no contribution, then

$$\hat{L}_2 \approx \hat{L}_1$$

In contrast, consider the value of the test statistic if the additional variable makes no contribution whatsoever to the risk of disease over and above that contributed by X_1 and X_2. This would mean that the maximized likelihood value $\hat{L}_2$ is essentially equal to the maximized likelihood value $\hat{L}_1$.

$$\Rightarrow \frac{\hat{L}_1}{\hat{L}_2} \approx 1$$

$$\Rightarrow \text{ LR } \approx -2\ln(1) = -2 \times 0 = 0$$

Thus, X_3 nonsignificant $\Rightarrow$ LR $\approx$ 0

Correspondingly, the ratio $\hat{L}_1$ divided by $\hat{L}_2$ is approximately equal to 1. Therefore, the likelihood ratio statistic is approximately equal to -2 times the natural log of 1, which is 0, because the log of 1 is 0. Thus, the likelihood ratio statistic for a highly nonsignificant X_3 variable is approximately 0. This, again, is what one would expect from a chi-square statistic.

$$\begin{array}{ccc} 0 \leq & \text{LR} & \leq \infty \\ \uparrow & & \uparrow \\ \text{N.S.} & & \text{S.} \end{array}$$

Similar to chi square (χ^2)

In summary, the likelihood ratio statistic, regardless of which two models are being compared, yields a value that lies between 0, when there is extreme nonsignificance, and $+\infty$, when there is extreme significance. This is the way a chi-square statistic works.

LR approximate χ^2 if n large

How large? No precise answer.

Statisticians have shown that the likelihood ratio statistic can be considered approximately chi square, provided that the number of subjects in the study is large. How large is large, however, has never been precisely documented, so the applied researcher has to have as large a study as possible and/or hope that the number of study subjects is large enough.

EXAMPLE

Model 2: logit $P_2(\mathbf{X}) = \alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$
(reduced model)

Model 3: logit $P_3(\mathbf{X}) = \alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_1X_3 + \beta_5X_2X_3$
(full model)

As another example of a likelihood ratio test, we consider a comparison of Model 2 with Model 3. Because Model 3 is larger than Model 2, we now refer to Model 3 as the full model and to Model 2 as the reduced model.

EXAMPLE (continued)

$H_0: \quad \beta_4 = \beta_5 = 0$
(similar to multiple–partial F test)

$H_A: \quad \beta_4$ and/or β_5 are not zero

There are two additional parameters in the full model that are not part of the reduced model; these are β_4 and β_5, the coefficients of the product variables X_1X_3 and X_2X_3, respectively. Thus, the null hypothesis that compares Models 2 and 3 is stated as β_4 equals β_5 equals 0. This is similar to the null hypothesis for a multiple-partial F test in classical multiple linear regression analysis. The alternative hypothesis here is that β_4 and/or β_5 are not 0.

$X_3 = E$

X_1, X_2 confounders

X_1X_3, X_2X_3 interaction terms

$H_0: \quad \beta_4 = \beta_5 = 0 \Leftrightarrow H_0:$ no interaction with E

If the variable X_3 is the exposure variable E in one's study and the variables X_1 and X_2 are confounders, then the product terms X_1X_3 and X_2X_3 are interaction terms for the interaction of E with X_1 and X_2, respectively. Thus, the null hypothesis that β_4 equals β_5 equals 0, is equivalent to testing no joint interaction of X_1 and X_2 with E.

$$\text{LR} = -2\ln\hat{L}_2 - \left(-2\ln\hat{L}_3\right) = -2\ln\left(\frac{\hat{L}_2}{\hat{L}_3}\right)$$

which is approximately χ^2 with 2 df under
$H_0: \quad \beta_4 = \beta_5 = 0$

The likelihood ratio statistic for comparing Models 2 and 3 is then given by $-2 \ln \hat{L}_2$ minus $-2 \ln \hat{L}_3$, which also can be written as -2 times the natural log of the ratio of $\hat{L}_2$ divided by $\hat{L}_3$. This statistic has an approximate chi-square distribution in large samples. The degrees of freedom here equals 2 because there are two parameters being set equal to 0 under the null hypothesis.

$$\underset{\uparrow}{-2 \ln \hat{L}_2}, \quad \underset{\uparrow}{-2 \ln \hat{L}_3}$$

Computer prints these
separately

When using a standard computer package to carry out this test, we must get the computer to fit the full and reduced models separately. The computer output for each model will include the log likelihood statistics of the form $-2 \ln \hat{L}$. The user then simply finds the two log likelihood statistics from the output for each model being compared and subtracts one from the other to get the likelihood ratio statistic of interest.

V. The Wald Test

Focus on 1 parameter

e.g., $H_0: \quad \beta_3 = 0$

There is another way to carry out hypothesis testing in logistic regression without using a likelihood ratio test. This second method is sometimes called *the Wald test*. This test is usually done when there is only one parameter being tested, as, for example, when comparing Models 1 and 2 above.

Wald statistic (for large n):

$Z = \frac{\hat{\beta}}{s_{\hat{\beta}}}$ is approximately $N(0,\,1)$

or

Z^2 is approximately χ^2 with 1 df

The Wald test statistic is computed by dividing the estimated coefficient of interest by its standard error. This test statistic has approximately a normal (0, 1), or Z, distribution in large samples. The square of this Z statistic is approximately a chi-square statistic with one degree of freedom.

Variable	ML Coefficient	S.E.	Chi sq	P
X_1	$\hat{\beta}_1$	$s_{\hat{\beta}_1}$	χ^2	P
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
X_{j}	$\hat{\beta}_j$	$s_{\hat{\beta}_j}$	χ^2	P
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
X_k	$\hat{\beta}_k$	$s_{\hat{\beta}_k}$	χ^2	P

In carrying out the Wald test, the information required is usually provided in the output, which lists each variable in the model followed by its ML coefficient and its standard error. Several packages also compute the chisquare statistic and a P-value.

When using the listed output, the user must find the row corresponding to the variable of interest and either compute the ratio of the estimated coefficient divided by its standard error or read off the chi-square statistic and its corresponding P-value from the output.

$\mathrm{LR} \approx Z^2_{\mathrm{Wald}}$ in large samples

The likelihood ratio statistic and its corresponding squared Wald statistic give approximately the same value in very large samples; so if one's study is large enough, it will not matter which statistic is used.

$\mathrm{LR} \neq Z^2_{\mathrm{Wald}}$ in small to moderate samples

LR preferred (statistical)

Wald convenient – fit only one model

Nevertheless, in small to moderate samples, the two statistics may give very different results. Statisticians have shown that the likelihood ratio statistic is better than the Wald statistic in such situations. So, when in doubt, it is recommended that the likelihood ratio statistic be used. However, the Wald statistic is somewhat convenient to use because only one model, the full model, needs to be fit.

EXAMPLE

Model 1: $\operatorname{logit} P_1(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2$

Model 2: $\operatorname{logit} P_2(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

$H_0: \quad \beta_3 = 0$

$Z = \frac{\hat{\beta}_3}{s_{\hat{\beta}_3}}$ is approximately $N(0,\,1)$

As an example of a Wald test, consider again the comparison of Models 1 and 2 described above. The Wald test for testing the null hypothesis that β_3 equals 0 is given by the Z statistic equal to $\hat{\beta}_3$ divided by the standard error of $\hat{\beta}_3$. The computed Z can be compared with percentage points from a standard normal table.

EXAMPLE (continued)

or
Z^2 is approximately χ^2 with 1 df

Or, alternatively, the Z can be squared and then compared with percentage points from a chi-square distribution with one degree of freedom.

Wald test for more than one parameter: requires matrices (See *Epidemiol. Res.*, Chap. 20, p. 431 for mathematical formula. Also, see Chapters 14 and 15 here for how used with correlated data.)

The Wald test we have just described considers a null hypothesis involving only one model parameter. There is also a generalized Wald test that considers null hypotheses involving more than one parameter, such as when comparing Models 2 and 3 above. However, the formula for this test requires knowledge of matrix theory and is beyond the scope of this presentation. The reader is referred to the text by Kleinbaum, Kupper, and Morgenstern (*Epidemiol. Res.*, Chap. 20, p. 431) for a description of this test. We refer to this test again in Chapters 14 and 15 when considering correlated data.

Third testing method:

Score statistic
(See Kleinbaum et al., *Commun. Stat.*, 1982 and Chapters 14 and 15 here.)

Yet another method for testing these hypotheses involves the use of a *score statistic* (see Kleinbaum et al., *Commun. Stat.*, 1982). Because this statistic is not routinely calculated by standard ML programs, and because its use gives about the same numerical chi-square values as the two techniques just presented, we will not discuss it further in this chaper.

VI. Interval Estimation: One Coefficient

Large sample confidence interval:

Estimate ± (percentage point of Z × estimated standard error)

We have completed our discussion of hypothesis testing and are now ready to describe *confidence interval estimation*. We first consider interval estimation when there is only one regression coefficient of interest. The procedure typically used is to obtain a large sample confidence interval for the parameter by computing *the estimate of the parameter plus or minus a percentage point of the normal distribution times the estimated standard error*.

EXAMPLE

Model 2: logit $P_2(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

100(1 − α)% CI for β_3: $\hat{\beta}_3 \pm Z_{1-\frac{\alpha}{2}} \times s_{\hat{\beta}_3}$

$\hat{\beta}_3$ and $s_{\hat{\beta}_3}$: from printout

Z from $N(0, 1)$ tables,

e.g., 95% ⇒ $\alpha = 0.05$

$$\Rightarrow 1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975$$

$Z_{0.975} = 1.96$

As an example, if we focus on the β_3 parameter in Model 2, the 100 times (1 − α)% confidence interval formula is given by $\hat{\beta}_3$ plus or minus the corresponding (1 − α/2)th percentage point of Z times the estimated standard error of $\hat{\beta}_3$.

In this formula, the values for $\hat{\beta}_3$ and its standard error are found from the printout. The Z percentage point is obtained from tables of the standard normal distribution. For example, if we want a 95% confidence interval, then α is 0.05, 1 − α/2 is 1 −0.025 or 0.975, and $Z_{0.975}$ is equal to 1.96.

CI for coefficient
vs.
✓ CI for odds ratio

Most epidemiologists are not interested in getting a confidence interval for the coefficient of a variable in a logistic model, but rather want a *confidence interval for an odds ratio* involving that parameter and possibly other parameters.

EXAMPLE

logit $P_2(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

$X_3 = (0, 1)$ variable

⇒ OR $= e^{\beta_3}$

CI for OR: exp(CI for β_3)

Model 2: $X_3 = (0, 1)$ exposure

X_1 and X_2 confounders

95% CI for OR:

$$\exp\left(\hat{\beta}_3 \pm 1.96 s_{\hat{\beta}_3}\right)$$

Above formula assumes X_3 is coded as (0, 1)

When only *one exposure variable*, is being considered, such as X_3 in Model 2, and this variable is a (0, 1) variable, then the odds ratio of interest, which adjusts for the other variables in the model, is e to that parameter, for example e to β_3. In this case, the corresponding *confidence interval for the odds ratio is obtained by exponentiating the confidence limits obtained for the parameter*.

Thus, if we consider Model 2, and if X_3 denotes a (0, 1) exposure variable of interest and X_1 and X_2 are confounders, then a 95% confidence interval for the adjusted odds ratio e to β_3 is given by the exponential of the confidence interval for β_3, as shown here.

This formula is correct, provided that the variable X_3 is a (0, 1) variable. If this variable is coded differently, such as (−1, 1), or if this variable is an ordinal or interval variable, then the confidence interval formula given here must be modified to reflect the coding.

Chapter 3: Computing OR for different codings

A detailed discussion of the effect of different codings of the exposure variable on the computation of the odds ratio is described in Chap. 3 of this text. It is beyond the scope of this presentation to describe in detail the effect of different codings on the corresponding confidence interval for the odds ratio. We do, however, provide a simple example to illustrate this situation.

EXAMPLE

$$X_3 \text{ coded as} \begin{cases} -1 & \text{unexposed} \\ 1 & \text{exposed} \end{cases}$$

$$\text{OR} = \exp[1 - (-1)\beta_3] = e^{2\beta_3}$$

$$95\% \text{ CI} : \exp\left(2\hat{\beta}_3 \pm 1.96 \times 2s_{\hat{\beta}_3}\right)$$

Suppose X_3 is coded as $(-1, 1)$ instead of $(0, 1)$, so that $-$ 1 denotes unexposed persons and 1 denotes exposed persons. Then, the odds ratio expression for the effect of X_3 is given by e to 1 minus -1 times β_3, which is e to 2 times β_3. The corresponding 95% confidence interval for the odds ratio is then given by exponentiating the confidence limits for the parameter $2\beta_3$, as shown here; that is, the previous confidence interval formula is modified by multiplying $\hat{\beta}_3$ and its standard error by the number 2.

VII. Interval Estimation: Interaction

No interaction: simple formula

Interaction: complex formula

The above confidence interval formulae involving a single parameter assume that there are no interaction effects in the model. When there is interaction, the confidence interval formula must be modified from what we have given so far. Because the general confidence interval formula is quite complex when there is interaction, our discussion of the modifications required will proceed by example.

EXAMPLE

Model 3: $X_3 = (0, 1)$ exposure

$$\text{logit } P_3(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3 + \beta_5 X_2 X_3$$

$$\widehat{\text{OR}} = \exp\left(\hat{\beta}_3 + \hat{\beta}_4 X_1 + \hat{\beta}_5 X_2\right)$$

Suppose we focus on Model 3, which is again shown here, and we assume that the variable X_3 is a (0, 1) exposure variable of interest. Then the formula for the estimated odds ratio for the effect of X_3 controlling for the variables X_1 and X_2 is given by the exponential of the quantity $\hat{\beta}_3$ plus $\hat{\beta}_4$ times X_1 plus $\hat{\beta}_5$ times X_2, where $\hat{\beta}_4$ and $\hat{\beta}_5$ are the estimated coefficients of the interaction terms X_1X_3 and X_2X_3 in the model.

EXAMPLE

i.e., $\widehat{\mathrm{OR}} = e^{\hat{l}}$,
where
$l = \beta_3 + \beta_4 X_1 + \beta_5 X_2$

We can alternatively write this estimated odds ratio formula as e to the $\hat{l}$, where l is the linear function β_3 plus β_4 times X_1 plus β_5 times X_2, and $\hat{l}$ is the estimate of this linear function using the ML estimates.

100 $(1-\alpha)$% CI for e^l
similar to CI formula for e^{β_3}

To obtain a 100 times $(1-\alpha)$% confidence interval for the odds ratio e to l, we must use the linear function l the same way that we used the single parameter β_3 to get a confidence interval for β_3. The corresponding confidence interval is thus given by exponentiating the confidence interval for l.

$$\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{var}}(\hat{l})}\right]$$

similar to $\exp\left[\hat{\beta}_3 \pm Z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{var}}(\hat{\beta}_3)}\right]$

$\sqrt{\widehat{\mathrm{var}}(\bullet)}$ = standard error

The formula is therefore the exponential of the quantity $\hat{l}$ plus or minus a percentage point of the Z distribution times the square root of the estimated variance of $\hat{l}$. Note that the square root of the estimated variance is the standard error.

General CI formula:

$$\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{var}}(\hat{l})}\right]$$

Example: $l = \beta_3 + \beta_4 X_1 + \beta_5 X_2$

This confidence interval formula, though motivated by our example using Model 3, is actually the general formula for the confidence interval for any odds ratio of interest from a logistic model. In our example, the linear function l took a specific form, but, in general, the linear function may take any form of interest.

General expression for l:

$$\mathrm{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} = e^{\sum\limits_{l=1}^{k} \beta_i (X_{1i} - X_{0i})}$$

OR $= e^l$ where

$$l = \sum_{i=1}^{k} \beta_i (X_{1i} - X_{0i})$$

A general expression for this linear function makes use of the general odds ratio formula described in our review. That is, the odds ratio comparing two groups identified by the vectors $\mathbf{X}_1$ and $\mathbf{X}_0$ is given by the formula e to the sum of terms of the form β_i times the difference between X_{1i} and X_{0i}, where the latter denotes the values of the ith variable in each group. We can equivalently write this as e to the l, where l is the linear function given by the sum of the β_i times the difference between X_{1i} and X_{0i}. This latter formula is the general expression for l.

Interaction: variance calculation difficult

No interaction: variance directly from printout

The difficult part in computing the confidence interval for an odds ratio involving interaction effects is the calculation for the estimated variance or corresponding square root, the standard error. When there is *no interaction*, so that the parameter of interest is a single regression coefficient, this variance is obtained directly from the variance–covariance output or from the listing of estimated coefficients and corresponding standard errors.

$$\text{var}(\hat{l}) = \text{var}\underbrace{\left[\sum\hat{\beta}_i(X_{1i} - X_{0i})\right]}_{\text{linear sum}}$$

$\hat{\beta}_i$ are correlated for different i

Must use var $\left(\hat{\beta}_i\right)$ and cov $\left(\hat{\beta}_i, \hat{\beta}_j\right)$

However, when the odds ratio involves *interaction* effects, the estimated variance considers a linear sum of estimated regression coefficients. The difficulty here is that, because the coefficients in the linear sum are estimated from the same data set, these coefficients are correlated with one another. Consequently, the calculation of the estimated variance must consider both the variances and the covariances of the estimated coefficients, which makes computations somewhat cumbersome.

EXAMPLE (model 3)

$$\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{var}}(\hat{l})}\right],$$

where $\hat{l} = \hat{\beta}_3 + \hat{\beta}_4 X_1 + \hat{\beta}_5 X_2$

Returning to the interaction example, recall that the confidence interval formula is given by exponentiating the quantity $\hat{l}$ plus or minus a Z percentage point times the square root of the estimated variance of $\hat{l}$, where $\hat{l}$ is given by $\hat{\beta}_3$ plus $\hat{\beta}_4$ times X_1 plus $\hat{\beta}_5$ times X_2.

$$\begin{aligned}\widehat{\text{var}}(\hat{l}) &= \widehat{\text{var}}\left(\hat{\beta}_3\right) + (X_1)^2\,\widehat{\text{var}}\left(\hat{\beta}_4\right)\\ &\quad + (X_2)^2\,\widehat{\text{var}}\left(\hat{\beta}_5\right)\\ &\quad + 2X_1\,\widehat{\text{cov}}\left(\hat{\beta}_3, \hat{\beta}_4\right)\\ &\quad + 2X_2\,\widehat{\text{cov}}\left(\hat{\beta}_3, \hat{\beta}_5\right)\\ &\quad + 2X_1X_2\,\widehat{\text{cov}}\left(\hat{\beta}_4, \hat{\beta}_5\right)\end{aligned}$$

It can be shown that the estimated variance of this linear function is given by the formula shown here.

var(β_i) and cov$\left(\hat{\beta}_i, \hat{\beta}_j\right)$ obtained from printout BUT must specify X_1 and X_2

The estimated variances and covariances in this formula are obtained from the estimated variance–covariance matrix provided by the computer output. However, the calculation of both $\hat{l}$ and the estimated variance of $\hat{l}$ requires additional specification of values for the effect modifiers in the model, which in this case are X_1 and X_2.

EXAMPLE (continued)

e.g., $X_1 = \text{AGE}$, $X_2 = \text{SMK}$:

Specification 1: $X_1 = 30$, $X_2 = 1$
versus
Specification 2: $X_1 = 40$, $X_2 = 0$

Different specifications yield different confidence intervals

For example, if X_1 denotes AGE and X_2 denotes smoking status (SMK), then one specification of these variables is $X_1 = 30$, $X_2 = 1$, and a second specification is $X_1 = 40$, $X_2 = 0$. Different specifications of these variables will yield different confidence intervals. This should be no surprise because a model containing interaction terms implies that both the estimated odds ratios and their corresponding confidence intervals vary as the values of the effect modifiers vary.

Recommendation. Use "typical" or "representative" values of X_1 and X_2 e.g., $\bar{X}_1$ and $\bar{X}_2$ in quintiles

A recommended practice is to use "typical" or "representative" values of X_1 and X_2, such as their mean values in the data, or the means of subgroups, for example, quintiles, of the data for each variable.

Some computer packages compute

$\widehat{\text{var}}(\hat{l})$

Some computer packages for logistic regression do compute the estimated variance of linear functions like $\hat{l}$ as part of the program options. See the Computer Appendix for details on the use of the "contrast" option in SAS and the "lincom" option in STATA.

General CI formula for *E, V, W* model:

$$\widehat{\text{OR}} = e^{\hat{l}},$$

where

$$l = \beta + \sum_{j=1}^{p_2} \delta_j W_j$$

$$\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{var}}(\hat{l})}\right],$$

where

$$\begin{aligned}\widehat{\text{var}}(\hat{l}) = \widehat{\text{var}}\left(\hat{\beta}\right) &+ \sum_{j=1}^{p_2} W_j^2\, \widehat{\text{var}}\left(\hat{\delta}_j\right) \\ &+ 2\sum_{j=1}^{p_2} W_j\, \widehat{\text{cov}}\left(\hat{\beta}, \hat{\delta}_j\right) \\ &+ 2\sum_j \sum_k W_j W_k\, \widehat{\text{cov}}\left(\hat{\delta}_j, \hat{\delta}_k\right)\end{aligned}$$

For the interested reader, we provide here the general formula for the estimated variance of the linear function obtained from the *E, V, W* model. Recall that the estimated odds ratio for this model can be written as e to $\hat{l}$, where l is the linear function given by the sum of β plus the sum of terms of the form δ_j times W_j.

The corresponding confidence interval formula is obtained by exponentiating the confidence interval for $\hat{l}$, where the variance of $\hat{l}$ is given by the general formula shown here.

Obtain $\widehat{\text{var}}$s and $\widehat{\text{cov}}$s from printout *but* must specify Ws.

In applying this formula, the user obtains the estimated variances and covariances from the variance–covariance output. However, as in the example above, the user must specify values of interest for the effect modifiers defined by the Ws in the model.

EXAMPLE

E, V, W model (Model 3):

$X_3 = E,$

$X_1 = V_1 = W_1$

$X_2 = V_2 = W_2$

$$\hat{l} = \hat{\beta}_3 + \hat{\beta}_4 X_1 + \hat{\beta}_5 X_2$$
$$= \hat{\beta} + \hat{\delta}_1 W_1 + \hat{\delta}_2 W_2$$

$\beta = \beta_3,$

$p_2 = 2,\ W_1 = X_1,\ W_2 = X_2,$
$\delta_1 = \beta_4$, and $\delta_2 = \beta_5$

Note that the example described earlier involving Model 3 is a special case of the formula for the *E, V, W* model, with X_3 equal to *E*, X_1 equal to both V_1 and W_1, and X_2 equal to both V_2 and W_2. The linear function *l* for Model 3 is shown here, both in its original form and in the *E, V, W* format.

To obtain the confidence interval for the Model 3 example from the general formula, the following substitutions would be made in the general variance formula: $\beta = \beta_3, p_2 = 2, W_1 = X_1, W_2 = X_2, \delta_1 = \beta_4$, and $\delta_2 = \beta_5$.

VIII. Numerical Example

EVANS COUNTY, GA
$n = 609$

Before concluding this presentation, we illustrate the ML techniques described above by way of a numerical example. We consider the printout results provided below and on the following page. These results summarize the computer output for two models based on follow-up study data on a cohort of 609 white males from Evans County, Georgia.

EXAMPLE

D = CHD (0, 1)

E = CAT

Cs = AGE, CHL, ECG, SMK, HPT
(conts) (conts) (0, 1) (0, 1) (0, 1)

Model A Output:
$-2 \ln \hat{L} = 400.39$

	Variable	Coefficient	S.E.	Chi sq	*P*
	Intercept	−6.7747	1.1402	35.30	0.0000
	CAT	0.5978	0.3520	2.88	0.0894
*V*s	AGE	0.0322	0.0152	4.51	0.0337
	CHL	0.0088	0.0033	7.19	0.0073
	ECG	0.3695	0.2936	1.58	0.2082
	SMK	0.8348	0.3052	7.48	0.0062
	HPT	0.4392	0.2908	2.28	0.1310

Unconditional ML estimation
$n = 609$, # parameters = 7

The outcome variable is coronary heart disease status, denoted as CHD, which is 1 if a person develops the disease and 0 if not. There are six independent variables of primary interest. The exposure variable is catecholamine level (CAT), which is 1 if high and 0 if low. The other independent variables are the control variables. These are denoted as AGE, CHL, ECG, SMK, and HPT.

The variable AGE is treated continuously. The variable CHL, which denotes cholesterol level, is also treated continuously. The other three variables are (0, 1) variables. ECG denotes electrocardiogram abnormality status, SMK denotes smoking status, and HPT denotes hypertension status.

EXAMPLE (continued)

Model A results are at bottom of previous page

The first set of results described by the printout information considers a model – called Model A – with no interaction terms. Thus, Model A contains the exposure variable CAT and the five covariables AGE, CHL, ECG, SMK, and HPT. Using the *E, V, W* formulation, this model contains five *V* variables, namely, the covariables, and no *W* variables.

Model B Output:
$-2 \ln \hat{L} = 347.23$

	Variable	Coefficient	S.E.	Chi sq	*P*
	Intercept	−4.0497	1.2550	10.41	0.0013
	CAT	−12.6894	3.1047	16.71	0.0000
*V*s	AGE	0.0350	0.0161	4.69	0.0303
	CHL	−0.0055	0.0042	1.70	0.1923
	ECG	0.3671	0.3278	1.25	0.2627
	SMK	0.7732	0.3273	5.58	0.0181
	HPT	1.0466	0.3316	9.96	0.0016
interaction	CH	−2.3318	0.7427	9.86	0.0017
	CC	0.0692	0.3316	23.20	0.0000

*W*s

CH = CAT HPT and CC = CAT CHL

unconditional ML estimation

$n = 609$, # parameters = 9

The second set of results considers Model B, which contains two interaction terms in addition to the variables contained in the first model. The two interaction terms are called CH and CC, where CH equals the product CAT × HPT and CC equals the product CAT × CHL. Thus, this model contains five *V* variables and two *W* variables, the latter being HPT and CHL.

Both sets of results have been obtained using unconditional ML estimation. Note that no matching has been done and that the number of parameters in each model is 7 and 9, respectively, which is quite small compared with the number of subjects in the data set, which is 609.

Model A: no interaction
$-2 \ln \hat{L} = 400.39$

Variable	Coefficient	S.E.	Chi sq	*P*
Intercept	−6.7747	1.1402	35.30	0.0000
CAT	0.5978	0.3520	2.88	0.0894
⋮			⋮	
HPT	0.4392	0.2908	2.28	0.1310

We focus for now on the set of results involving the no interaction Model A. The information provided consists of the log likelihood statistic $-2 \ln \hat{L}$ at the top followed by a listing of each variable and its corresponding estimated coefficient, standard error, chi-square statistic, and *P*-value.

$\widehat{\text{OR}} = \exp(0.5978) = 1.82$

For this model, because CAT is the exposure variable and there are no interaction terms, the estimated odds ratio is given by e to the estimated coefficient of CAT, which is e to the quantity 0.5978, which is 1.82. Because Model A contains five *V* variables, we can interpret this odds ratio as an adjusted odds ratio for the effect of the CAT variable, which controls for the potential confounding effects of the five *V* variables.

Test statistic	Info. available?
LR	No
Wald	Yes

We can use this information to carry out a hypothesis test for the significance of the estimated odds ratio from this model. Of the two test procedures described, namely, the likelihood ratio test and the Wald test, the information provided only allows us to carry out the Wald test.

EXAMPLE (continued)

LR test:

Full model	Reduced model
Model A	Model A w/o CAT

H_0: $\beta = 0$

where β = coefficient of CAT in model A

Reduced model (w/o CAT) printout not provided here

WALD TEST:

Variable	Coefficient	S.E.	Chi sq	*P*
Intercept	–6.7747	1.1402	35.30	0.0000
CAT	0.5978	0.3520	2.88	0.0894
AGE	0.0322	0.0152	4.51	0.0337
CHL	0.0088	0.0033	7.19	0.0073
ECG	0.3695	0.2936	1.58	0.2082
SMK	0.8348	0.3052	7.48	0.0062
HPT	0.4392	0.2908	2.28	0.1310

$$Z = \frac{0.5978}{0.3520} = 1.70$$

$$Z^2 = \text{CHISQ} = \boxed{2.88}$$

$P = 0.0896$ misleading
(Assumes two-tailed test)
usual question: OR > 1? (one-tailed)

$$\text{One-tailed}\,P = \frac{\text{Two-tailed}\,P}{2} = \frac{0.0894}{2} = \boxed{0.0447}$$

$P < 0.05 \Rightarrow$ significant at 5% level

To carry out the *likelihood ratio test*, we would need to compare two models. The full model is Model A as described by the first set of results discussed here. The reduced model is a different model that contains the five covariables without the CAT variable.

The null hypothesis here is that the coefficient of the CAT variable is zero in the full model. Under this null hypothesis, the model will reduce to a model without the CAT variable in it. Because we have provided neither a printout for this reduced model nor the corresponding log likelihood statistic, we cannot carry out the likelihood ratio test here.

To carry out the *Wald test* for the significance of the CAT variable, we must use the information in the row of results provided for the CAT variable. The Wald statistic is given by the estimated coefficient divided by its standard error; from the results, the estimated coefficient is 0.5978 and the standard error is 0.3520.

Dividing the first by the second gives us the value of the Wald statistic, which is a *Z*, equal to 1.70. Squaring this statistic, we get the chi-square statistic equal to 2.88, as shown in the table of results.

The *P*-value of 0.0894 provided next to this chi square is somewhat misleading. This *P*-value considers a two-tailed alternative hypothesis, whereas most epidemiologists are interested in one-tailed hypotheses when testing for the significance of an exposure variable. That is, the usual question of interest is whether the odds ratio describing the effect of CAT controlling for the other variables is significantly *higher* than the null value of 1.

To obtain a one-tailed *P*-value from a two-tailed *P*-value, we simply take half of the two-tailed *P*-value. Thus, for our example, the one-tailed *P*-value is given by 0.0894 divided by 2, which is 0.0447. Because this *P*-value is less than 0.05, we can conclude, assuming this model is appropriate, that there is a significant effect of the CAT variable at the 5% level of significance.

EXAMPLE (continued)

H_0: $\beta = 0$
equivalent to
H_0: adjusted OR = 1

The Wald test we have just described tests the null hypothesis that the coefficient of the CAT variable is 0 in the model containing CAT and five covariables. An equivalent way to state this null hypothesis is that the odds ratio for the effect of CAT on CHD adjusted for the five covariables is equal to the null value of 1.

Variable	Coefficient	S.E.	Chi sq	*P*
Intercept				
CAT				
AGE				
CHL	0.0088	0.0033	7.18	0.0074
⋮			↑	
HPT			Not of interest	

The other chi-square statistics listed in the table provide Wald tests for other variables in the model. For example, the chi-square value for the variable CHL is the squared Wald statistic that tests whether there is a significant *effect of CHL* on CHD controlling for the other five variables listed, including CAT. However, the Wald test for CHL, or for any of the other five covariables, is not of interest in this study because the only exposure variable is CAT and because the other five variables are in the model for control purposes.

95% CI for adjusted OR:
First, 95% CI for β:

$\hat{\beta} \pm 1.96 \times s_{\hat{\beta}}$

$0.5978 \pm 1.96 \times 0.3520$

CI limits for β: $(-0.09, 1.29)$

A 95% confidence interval for the odds ratio for the adjusted effect of the CAT variable can be computed from the set of results for the no interaction model as follows: We first obtain a confidence interval for β, the coefficient of the CAT variable, by using the formula $\hat{\beta}$ plus or minus 1.96 times the standard error of $\hat{\beta}$. This is computed as 0.5978 plus or minus 1.96 times 0.3520. The resulting confidence limits for $\hat{\beta}$ are -0.09 for the lower limit and 1.29 for the upper limit.

exp(CI limits for β) $= (e^{-0.09}, e^{1.29})$

$\boxed{= (0.91, 3.63)}$

Exponentiating the lower and upper limits gives the confidence interval for the adjusted odds ratio, which is 0.91 for the lower limit and 3.63 for the upper limit.

CI contains 1,
so
do not reject H_0
at
5% level (*two-tailed*)

Note that this confidence interval contains the value 1, which indicates that a two-tailed test is not significant at the 5% level statistical significance from the Wald test. This does not contradict the earlier Wald test results, which were significant at the 5% level because using the CI, our alternative hypothesis is two-tailed instead of one-tailed.

EXAMPLE (continued)

No interaction model
vs.
other models?

Note that the no interaction model we have been focusing on may, in fact, be inappropriate when we compare it to other models of interest. In particular, we now compare the no interaction model to the model described by the second set of printout results we have provided.

Model B vs. Model A

We will see that this second model, B, which involves interaction terms, is a better model. Consequently, the results and interpretations made about the effect of the CAT variable from the no interaction Model A may be misleading.

LR test for interaction:
$H_0: \quad \delta_1 = \delta_2 = 0$

where δs are coefficients of interaction terms CC and CH in model B

To compare the no interaction model with the interaction model, we need to carry out a *likelihood ratio test for the significance of the interaction terms*. The null hypothesis here is that the coefficients δ_1 and δ_2 of the two interaction terms are both equal to 0.

Full Model	Reduced Model
Model B (interaction)	Model A (no interaction)

For this test, the full model is the interaction Model B and the reduced model is the no interaction Model A. The likelihood ratio test statistic is then computed by taking the difference between log likelihood statistics for the two models.

$$\begin{aligned} \text{LR} &= -2\ln \hat{L}_{\text{model A}} - (-2\ln \hat{L}_{\text{model B}}) \\ &= 400.39 - 347.23 \\ &= 53.16 \end{aligned}$$

df = 2
significant at .01 level

From the printout information given on pages 146–147, this difference is given by 400.39 minus 347.23, which equals 53.16. The degrees of freedom for this test is 2 because there are two parameters being set equal to 0. The chi-square statistic of 53.16 is found to be significant at the: 01 level. Thus, the likelihood ratio test indicates that the interaction model is better than the no interaction model.

$\widehat{\text{OR}}$ for interaction model (B):

$$\widehat{\text{OR}} = \exp\left(\hat{\beta} + \hat{\delta}_1\text{CHL} + \hat{\delta}_2\text{HPT}\right)$$

$\hat{\beta} = -12.6894$ for CAT

$\hat{\delta}_1 = \ 0.0692$ for CC

$\hat{\delta}_2 = -2.3318$ for CH

We now consider what the odds ratio is for the interaction model. As this model contains product terms CC and CH, where CC is CAT × CHL and CH is CAT × HPT, the estimated odds ratio for the effect of CAT must consider the coefficients of these terms as well as the coefficient of CAT. The formula for this estimated odds ratio is given by the exponential of the quantity $\hat{\beta}$ plus δ_1 times CHL plus $\hat{\delta}_2$ times HPT, where $\hat{\beta}$ (−12.6894) is the coefficient of CAT, $\hat{\delta}_1$(0.0692) is the coefficient of the interaction term CC, and $\hat{\delta}_2$ (−2.3318) is the coefficient of the interaction term CH.

EXAMPLE (continued)

$$\widehat{\text{OR}} = \exp[\beta + \delta_1\text{CHL} + \delta_2\text{ HPT}]$$
$$= \exp[-12.6894 + 0.0692\text{ CHL} + (-2.3318)\text{HPT}]$$

Plugging the estimated coefficients into the odds ratio formula yields the expression: e to the quantity −12.6894 plus 0.0692 times CHL plus −2.3318 times HPT.

Must specify
CHL and HPT
↑ ↑
Effect modifiers

To obtain a numerical value from this expression, it is necessary to specify a value for CHL and a value for HPT. Different values for CHL and HPT will, therefore, yield different odds ratio values. This should be expected because the model with interaction terms should give different odds ratio estimates depending on the values of the effect modifiers, which in this case are CHL and HPT.

Adjusted $\widehat{\text{OR}}$:

		HPT	
		0	1
	200	3.16	0.31
CHL	220	12.61	1.22
	240	50.33	4.89

The table shown here illustrates different odds ratio estimates that can result from specifying different values of the effect modifiers. In this table, the values of CHL used are 200, 220, and 240; the values of HPT are 0 and 1. The cells within the table give the estimated odds ratios computed from the above expression for the odds ratio for different combinations of CHL and HPT.

CHL = 200, HPT = 0 ⇒ $\widehat{\text{OR}} = 3.16$

CHL = 220, HPT = 1 ⇒ $\widehat{\text{OR}} = 1.22$

$\widehat{\text{OR}}$ adjusts for AGE, CHL, ECG, SMK, and HPT (*V* variables)

For example, when CHL equals 200 and HPT equals 0, the estimated odds ratio is given by 3.16; when CHL equals 220 and HPT equals 1, the estimated odds ratio is 1.22. Each of the estimated odds ratios in this table describes the association between CAT and CHD adjusted for the five covariables AGE, CHL, ECG, SMK, and HPT because each of the covariables is contained in the model as *V* variables.

Confidence intervals:

$$\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{var}}(\hat{l})}\right]$$

where

$$\hat{l} = \hat{\beta} + \sum_{j=1}^{p_2} \hat{\delta}_j W_j$$

To account for the variability associated with each of the odds ratios presented in the above tables, we can compute confidence intervals by using the methods we have described. The general confidence interval formula is given by e to the quantity $\hat{l}$ plus or minus a percentage point of the *Z* distribution times the square root of the estimated variance of $\hat{l}$, where *l* is the linear function shown here.

EXAMPLE

$$\begin{aligned}\widehat{\text{var}}(\hat{l}) &= \widehat{\text{var}}\left(\hat{\beta}\right) + (W_1)^2\,\widehat{\text{var}}\left(\hat{\delta}_1\right)\\ &\quad + (W_2)^2\,\widehat{\text{var}}\left(\hat{\delta}_2\right) + 2W_1\,\widehat{\text{Cov}}\left(\hat{\beta},\hat{\delta}_1\right)\\ &\quad + 2W_2\,\widehat{\text{Cov}}\left(\hat{\beta},\hat{\delta}_2\right)\\ &\quad + 2W_1W_2\,\widehat{\text{Cov}}\left(\hat{\delta}_1,\hat{\delta}_2\right)\end{aligned}$$

$W_1 = \text{CHL}$, $W_2 = \text{HPT}$

$\text{CHL} = 220$, $\text{HPT} = 1$:

$$\begin{aligned}\hat{l} &= \hat{\beta} + \hat{\delta}_1(220) + \hat{\delta}_2(1)\\ &= 0.1960\end{aligned}$$

$$\widehat{\text{var}}(\hat{l}) = 0.2279$$

$$\hat{V} = \begin{bmatrix} \widehat{\text{var}}\hat{\beta} & & \\ \widehat{\text{cov}}\left(\hat{\beta},\hat{\delta}_1\right) & \widehat{\text{var}}\hat{\delta}_1 & \\ \widehat{\text{cov}}\left(\hat{\beta},\hat{\delta}_2\right) & \widehat{\text{cov}}\left(\hat{\delta}_1,\hat{\delta}_2\right) & \widehat{\text{var}}\hat{\delta}_2 \end{bmatrix}$$

$$= \begin{bmatrix} 9.6389 & & \\ -0.0437 & 0.0002 & \\ -0.0049 & -0.0016 & 0.5516 \end{bmatrix}$$

95% CI for adjusted OR:

$$\exp\left[0.1960 \pm 1.96\sqrt{0.2279}\right]$$

CI limits: (0.48, 3.10)

	HPT = 0	HPT = 1
CHL = 200	$\widehat{\text{OR}}$: 3.16 CI : (0.89, 11.03)	$\widehat{\text{OR}}$: 0.31 CI : (0.10, 0.91)
CHL = 220	$\widehat{\text{OR}}$: 12.61 CI : (3.65, 42.94)	$\widehat{\text{OR}}$: 1.22 CI : (0.48, 3.10)
CHL = 240	$\widehat{\text{OR}}$: 50.33 CI : (11.79, 212.23)	$\widehat{\text{OR}}$: 4.89 CI : (1.62, 14.52)

For the specific interaction model (B) we have been considering, the variance of $\hat{l}$ is given by the formula shown here.

In computing this variance, there is an issue concerning round-off error. Computer packages typically maintain 16 decimal places for calculations, with final answers rounded to a set number of decimals (e.g., 4) on the printout. The variance results we show here were obtained with such a program (see Computer Appendix) rather than using the rounded values from the variance–covariance matrix presented at left.

For this model, W_1 is CHL and W_2 is HPT.

As an example of a confidence interval calculation, we consider the values CHL equal to 220 and HPT equal to 1. Substituting $\hat{\beta}$, $\hat{\delta}_1$, and $\hat{\delta}_2$ into the formula for $\hat{l}$, we obtain the estimate $\hat{l}$ equals 0.1960.

The corresponding estimated variance is obtained by substituting into the above variance formula the estimated variances and covariances from the variance–covariance matrix $\hat{V}$. The resulting estimate of the variance of $\hat{l}$ is equal to 0.2279. The numerical values used in this calculation are shown at left.

We can combine the estimates of $\hat{l}$ and its variance to obtain the 95% confidence interval. This is given by exponentiating the quantity 0.1960 plus or minus 1.96 times the square root of 0.2279. The resulting confidence limits are 0.48 for the lower limit and 3.10 for the upper limit.

The 95% confidence intervals obtained for other combinations of CHL and HPT are shown here. For example, when CHL equals 200 and HPT equals 1, the confidence limits are 0.10 and 0.91. When CHL equals 240 and HPT equals 1, the limits are 1.62 and 14.52.

EXAMPLE

Wide CIs ⇒ estimates have large variances

HPT = 1:

$\widehat{\text{OR}} = 0.31$, CI: (0.10, .91) below 1

$\widehat{\text{OR}} = 1.22$, CI: (0.48, 3.10) includes 1

$\widehat{\text{OR}} = 4.89$, CI: (1.62, 14.52) above 1

	$\widehat{\text{OR}}$	(Two-tailed) significant?
CHL = 200:	0.31	Yes
220:	1.22	No
240:	4.89	Yes

All confidence intervals are quite wide, indicating that their corresponding point estimates have large variances. Moreover, if we focus on the three confidence intervals corresponding to HPT equal to 1, we find that the interval corresponding to the estimated odds ratio of 0.31 lies completely below the null value of 1. In contrast, the interval corresponding to the estimated odds ratio of 1.22 surrounds the null value of 1, and the interval corresponding to 4.89 lies completely above 1.

From a hypothesis testing standpoint, these results therefore indicate that the estimate of 1.22 is not statistically significant at the 5% level, whereas the other two estimates are statistically significant at the 5% level.

SUMMARY

Chapter 5: Statistical Inferences Using ML Techniques

This presentation is now complete. In summary, we have described two test procedures, the likelihood ratio test and the Wald test. We have also shown how to obtain interval estimates for odds ratios obtained from a logistic regression. In particular, we have described confidence interval formula for models with and without interaction terms.

We suggest that the reader review the material covered here by reading the summary outline that follows. Then you may work the practice exercises and test.

Chapter 6: Modeling Strategy Guidelines

In the next chapter, "Modeling Strategy Guidelines", we provide guidelines for determining a best model for an exposure–disease relationship that adjusts for the potential confounding and effect-modifying effects of covariables.

Detailed Outline

I. Overview (page 132)

Focus:

- Testing hypotheses
- Computing confidence intervals

II. Information for making statistical inferences (pages 132–133)

A. Maximized likelihood value: $L(\hat{\boldsymbol{\theta}})$.

B. Estimated variance–covariance matrix: $\hat{V}(\hat{\boldsymbol{\theta}})$ contains variances of estimated coefficients on the diagonal and covariances between coefficients off the diagonal.

C. Variable listing: contains each variable followed by ML estimate, standard error, and other information.

III. Models for inference-making (pages 133–134)

A. Model 1: logit $P(\mathbf{X}) = \alpha + \beta_1X_1 + \beta_2X_2$;
Model 2: logit $P(\mathbf{X}) = \alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$;
Model 3: logit $P(\mathbf{X}) = \alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_1X_3 + \beta_5X_2X_3$.

B. $\hat{L}_1, \hat{L}_2, \hat{L}_3$ are maximized likelihoods ($\hat{L}$) for models 1–3, respectively.

C. $\hat{L}$ is similar to R square: $\hat{L}_1 \leq \hat{L}_2 \leq \hat{L}_3$.

D. $-2 \ln \hat{L}_3 \leq -2 \ln \hat{L}_2 \leq -2 \ln \hat{L}_1$,
where $-2 \ln \hat{L}$ is called the log likelihood statistic.

IV. The likelihood ratio (LR) test (pages 134–138)

A. LR statistic compares two models: full (larger) model vs. reduced (smaller) model.

B. H_0: some parameters in full model are equal to 0.

C. df = number of parameters in full model set equal to 0 to obtain reduced model.

D. Model 1 vs. Model 2: LR $= -2 \ln \hat{L}_1 - (-2 \ln \hat{L}_2)$, where H_0: $\beta_3 = 0$. This LR has approximately a chi-square distribution with one df under the null hypothesis.

E. $-2\ln\hat{L}_1 - \left(-2\ln\hat{L}_2\right) = -2\ln\left(\hat{L}_1/\hat{L}_2\right)$,
where $\hat{L}_1/\hat{L}_2$ is a ratio of likelihoods.

F. How the LR test works: LR works like a chi-square statistic. For highly significant variables, LR is large and positive; for nonsignificant variables, LR is close to 0.

G. Model 2 vs. Model 3: LR $= -2 \ln \hat{L}_2 - (-2 \ln \hat{L}_3)$, where H_0: $\beta_4 = \beta_5 = 0$. This LR has approximately a chi-square distribution with 2 df under the null hypothesis.

H. Computer prints $-2 \ln \hat{L}$ separately for each model, so LR test requires only subtraction.

V. The Wald test (pages 138–140)

A. Requires one parameter only to be tested, e.g., H_0: $\beta_3 = 0$.

B. Test statistic: $Z = \hat{\beta}/s_{\hat{\beta}}$ which is approximately $N(0, 1)$ under H_0.

C. Alternatively, Z^2 is approximately chi square with one df under H_0.

D. LR and Z are approximately equal in large samples, but may differ in small samples.

E. LR is preferred for statistical reasons, although Z is more convenient to compute.

F. Example of Wald statistic for H_0: $\beta_3 = 0$ in Model 2: $Z = \hat{\beta}_3/s_{\hat{\beta}_3}$.

VI. Interval estimation: one coefficient (pages 140–142)

A. Large sample confidence interval: estimate ± percentage point of Z × estimated standard error.

B. 95% CI for β_3 in Model 2: $\hat{\beta}_3 \pm 1.96 s_{\hat{\beta}_3}$.

C. If X_3 is a (0, 1) exposure variable in Model 2, then the 95% CI for the odds ratio of the effect of exposure adjusted for X_1 and X_2 is given by $\exp\left(\hat{\beta}_3 \pm 1.96 s_{\hat{\beta}_3}\right)$

D. If X_3 has coding other than (0, 1), the CI formula must be modified.

VII. Interval estimation: interaction (pages 142–146)

A. Model 3 example: $\widehat{\text{OR}} = e^{\hat{l}}$, where $\hat{l} = \hat{\beta}_3 + \hat{\beta}_4 X_1 + \hat{\beta}_5 X_2$

100(1 −α)% CI formula for OR: $\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{var}}(\hat{l})}\right]$,

where

$$\widehat{\text{var}}(\hat{l}) = \widehat{\text{var}}\left(\hat{\beta}_3\right) + (X_1)^2\,\widehat{\text{var}}\left(\hat{\beta}_4\right) + (X_2)^2\,\widehat{\text{var}}\left(\hat{\beta}_5\right) + 2X_1\,\widehat{\text{cov}}\left(\hat{\beta}_3, \hat{\beta}_4\right) + 2X_2\,\widehat{\text{cov}}\left(\hat{\beta}_3, \hat{\beta}_5\right) + 2X_1X_2\,\widehat{\text{cov}}\left(\hat{\beta}_4, \hat{\beta}_5\right).$$

B. General 100(1 − α)% CI formula for OR:

$\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{var}}(\hat{l})}\right]$

where $\widehat{\text{OR}} = e^{\hat{l}}$,

$$\hat{l} = \sum_{i=1}^{k} \hat{\beta}_i (X_{1i} - X_{0i}) \text{ and } \text{var}(\hat{l}) = \text{var}\left(\underbrace{\Sigma\hat{\beta}_i(X_{1i} - X_{0i})}_{\text{linear sum}}\right).$$

C. 100(1 − α)% CI formula for OR using *E, V, W* model:

$$\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\text{var}}(\hat{l})}\right],$$

where $\widehat{\text{OR}} = e^{\hat{l}}, \hat{l} = \hat{\beta} + \sum_{j=1}^{p_2} \hat{\delta}_j W_j$

and $$\widehat{\text{var}}(\hat{l}) = \widehat{\text{var}}\left(\hat{\beta}\right) + \sum_{j=1}^{p_2} W_j^2\,\widehat{\text{var}}\left(\hat{\delta}_j\right)$$
$$+2\sum_{j=1}^{p_2} W_j\,\widehat{\text{cov}}\left(\hat{\beta}, \hat{\delta}_j\right)$$
$$+2\sum_j\sum_k W_j W_k\,\widehat{\text{cov}}\left(\hat{\delta}_j, \hat{\delta}_k\right).$$

D. Model 3 example of *E, V, W* model: $X_3 = E$, $X_1 = V_1, X_2 = V_2$, and for interaction terms, $p_2 = 2, X_1 = W_1, X_2 = W_2$.

VIII. Numerical example (pages 146–153)

A. Printout provided for two models (A and B) from Evans County, Georgia data.

B. Model A: no interaction terms; Model B: interaction terms.

C. Description of LR and Wald tests for Model A.

D. LR test for no interaction effect in Model B: compares model B (full model) with Model A (reduced model). Result: significant interaction.

E. 95% CI for OR from Model B; requires use of CI formula for interaction, where $p_2 = 2$, $W_1 = \text{CHL}$, and $W_2 = \text{HPT}$.

Practice Exercises

A prevalence study of predictors of surgical wound infection in 265 hospitals throughout Australia collected data on 12,742 surgical patients (McLaws et al., 1988). For each patient, the following independent variables were determined: type of hospital (public or private), size of hospital (large or small), degree of contamination of surgical site (clean or contaminated), and age and sex of the patient. A logistic model was fit to this data to predict whether or not the patient developed a surgical wound infection during hospitalization. The largest model fit included all of the above variables and all possible two-way interaction terms. The abbreviated variable names and the manner in which the variables were coded in the model are described as follows:

Variable	Abbreviation	Coding
Type of hospital	HT	1 = public, 0 = private
Size of hospital	HS	1 = large, 0 = small
Degree of contamination	CT	1 = contaminated, 0 = clean
Age	AGE	continuous
Sex	SEX	1 = female, 0 = male

1. State the logit form of a no interaction model that includes all of the above predictor variables.
2. State the logit form of a model that extends the model of Exercise 1 by adding all possible pairwise products of different variables.
3. Suppose you want to carry out a (global) test for whether any of the two-way product terms (considered collectively) in your interaction model of Exercise 2 are significant. State the null hypothesis, the form of the appropriate (likelihood ratio) test statistic, and the distribution and degrees of freedom of the test statistic under the null hypothesis of no interaction effects in your model of Exercise 2.

Suppose the test for interaction in Exercise 3 is nonsignificant, so that you felt justified to drop all pairwise products from your model. The remaining model will, therefore, contain only those variables given in the above listing.

4. Consider a test for the effect of hospital type (HT) adjusted for the other variables in the no interaction model. Describe the likelihood ratio test for this effect by stating the following: the null hypothesis, the formula for the test statistic, and the distribution and degrees of freedom of the test statistic under the null hypothesis.
5. For the same question as described in Exercise 4, that is, concerning the effect of HT controlling for the other variables in the model, describe the Wald test for this effect by providing the null hypothesis, the formula for the test statistic, and the distribution of the test statistic under the null hypothesis.
6. Based on the study description preceding Exercise 1, do you think that the likelihood ratio and Wald test results will be approximately the same? Explain.
7. Give a formula for a 95% confidence interval for the odds ratio describing the effect of HT controlling for the other variables in the no interaction model.

(*Note*. In answering all of the above questions, make sure to state your answers in terms of the coefficients and variables that you specified in your answers to Exercises 1 and 2).

Consider the following printout results that summarize the computer output for two models based on follow-up study data on 609 white males from Evans County, Georgia:

Model I OUTPUT:
$-2 \ln \hat{L} = 400.39$

Variable	Coefficient	S.E.	Chi sq	*P*
Intercept	−6.7747	1.1402	35.30	0.0000
CAT	0.5978	0.3520	2.88	0.0894
AGE	0.0322	0.0152	4.51	0.0337
CHL	0.0088	0.0033	7.19	0.0073
ECG	0.3695	0.2936	1.58	0.2082
SMK	0.8348	0.3052	7.48	0.0062
HPT	0.4392	0.2908	2.28	0.1310

Model II OUTPUT:
$-2 \ln \hat{L} = 357.05$

Variable	Coefficient	S.E.	Chi sq	*P*
Intercept	−3.9346	1.2503	9.90	0.0016
CAT	−14.0809	3.1227	20.33	0.0000
AGE	0.0323	0.0162	3.96	0.0466
CHL	−0.0045	0.00413	1.16	0.2821
ECG	0.3577	0.3263	1.20	0.2729
SMK	0.8069	0.3265	6.11	0.0134
HPT	0.6069	0.3025	4.03	0.0448
CC = CAT × CHL	0.0683	0.0143	22.75	0.0000

In the above models, the variables are coded as follows: CAT(1 = high, 0 = low), AGE(continuous), CHL(continuous), ECG(1 = abnormal, 0 = normal), SMK(1 = ever, 0 = never), HPT(1 = hypertensive, 0 = normal). The outcome variable is CHD status(1 = CHD, 0 = no CHD).

8. For Model I, test the hypothesis for the effect of CAT on the development of CHD. State the null hypothesis in terms of an odds ratio parameter, give the formula for the test statistic, state the distribution of the test statistic under the null hypothesis, and, finally, carry out the test for a one-sided alternative hypothesis using the above printout for Model I. Is the test significant?

9. Using the printout for Model I, compute the point estimate and a 95% confidence interval for the odds ratio for the effect of CAT on CHD controlling for the other variables in the model.
10. Now consider Model II: Carry out the likelihood ratio test for the effect of the product term CC on the outcome, controlling for the other variables in the model. Make sure to state the null hypothesis in terms of a model coefficient, give the formula for the test statistic and its distribution and degrees of freedom under the null hypothesis, and report the *P*-value. Is the test result significant?
11. Carry out the Wald test for the effect of CC on outcome, controlling for the other variables in Model II. In carrying out this test, provide the same information as requested in Exercise 10. Is the test result significant? How does it compare to your results in Exercise 10? Based on your results, which model is more appropriate, Model I or II?
12. Using the output for Model II, give a formula for the point estimate of the odds ratio for the effect of CAT on CHD, which adjusts for the confounding effects of AGE, CHL, ECG, SMK, and HPT and allows for the interaction of CAT with CHL.
13. Use the formula for the adjusted odds ratio in Exercise 12 to compute numerical values for the estimated odds ratio for the following cholesterol values: CHL = 220 and CHL = 240.
14. Give a formula for the 95% confidence interval for the adjusted odds ratio described in Exercise 12 when CHL = 220. In stating this formula, make sure to give an expression for the estimated variance portion of the formula in terms of variances and covariances obtained from the variance–covariance matrix.

Test

The following printout provides information for the fitting of two logistic models based on data obtained from a matched case-control study of cervical cancer in 313 women from Sydney, Australia (Brock et al., 1988). The outcome variable is cervical cancer status (1 = present, 0 = absent). The matching variables are age and socioeconomic status. Additional independent variables not matched on are smoking status, number of lifetime sexual partners, and age at first sexual intercourse. The independent variables not involved in the matching are listed below, together with their computer abbreviation and coding scheme.

Variable	Abbreviation	Coding
Smoking status	SMK	1 = ever, 0 = never
Number of sexual partners	NS	1 = 4+, 0 = 0–3
Age at first intercourse	AS	1 = 20+, 0 = ≤ 19

PRINTOUT:

Model I

$-2 \ln \hat{L} = 174.97$

Variable	β	S.E.	Chi sq	P
SMK	1.4361	0.3167	20.56	0.0000
NS	0.9598	0.3057	9.86	0.0017
AS	−0.6064	0.3341	3.29	0.0695

Model II

$-2 \ln \hat{L} = 171.46$

Variable	β	S.E.	Chi sq	P
SMK	1.9381	0.4312	20.20	0.0000
NS	1.4963	0.4372	11.71	0.0006
AS	−0.6811	0.3473	3.85	0.0499
SMK×NS	−1.1128	0.5997	3.44	0.0635

Variance–Covariance Matrix (Model II)

	SMK	NS	AS	SMK × NS
SMK	0.1859			
NS	0.1008	0.1911		
AS	−0.0026	−0.0069	0.1206	
SMK × NS	−0.1746	−0.1857	0.0287	0.3596

1. What method of estimation was used to obtain estimates of parameters for both models, conditional or unconditional ML estimation? Explain.
2. Why are the variables, age and socioeconomic status, missing from the printout, even though these were variables matched on in the study design?
3. For Model I, test the hypothesis for the effect of SMK on cervical cancer status. State the null hypothesis in

terms of an odds ratio parameter, give the formula for the test statistic, state the distribution of the test statistic under the null hypothesis, and, finally, carry out the test using the above printout for Model I. Is the test significant?

4. Using the printout for Model I, compute the point estimate and 95% confidence interval for the odds ratio for the effect of SMK controlling for the other variables in the model.
5. Now consider Model II: Carry out the likelihood ratio test for the effect of the product term SMK × NS on the outcome, controlling for the other variables in the model. Make sure to state the null hypothesis in terms of a model coefficient, give the formula for the test statistic and its distribution and degrees of freedom under the null hypothesis, and report the *P*-value. Is the test significant?
6. Carry out the Wald test for the effect of SMK × NS, controlling for the other variables in Model II. In carrying out this test, provide the same information as requested in Question 3. Is the test significant? How does it compare to your results in Question 5?
7. Using the output for Model II, give a formula for the point estimate of the odds ratio for the effect of SMK on cervical cancer status, which adjusts for the confounding effects of NS and AS and allows for the interaction of NS with SMK.
8. Use the formula for the adjusted odds ratio in Question 7 to compute numerical values for the estimated odds ratios when NS = 1 and when NS = 0.
9. Give a formula for the 95% confidence interval for the adjusted odds ratio described in Question 8 (when NS = 1). In stating this formula, make sure to give an expression for the estimated variance portion of the formula in terms of variances and covariances obtained from the variance–covariance matrix.
10. Use your answer to Question 9 and the estimated variance–covariance matrix to carry out the computation of the 95% confidence interval described in Question 7.
11. Based on your answers to the above questions, which model, point estimate, and confidence interval for the effect of SMK on cervical cancer status are more appropriate, those computed for Model I or those computed for Model II? Explain.

Answers to Practice Exercises

1. $\text{logit P}(\mathbf{X}) = \alpha + \beta_1\text{HT} + \beta_2\text{HS} + \beta_3\text{CT} + \beta_4\text{AGE} + \beta_5\text{SEX}.$
2. $\text{logit P}(\mathbf{X}) = \alpha + \beta_1\text{HT} + \beta_2\text{HS} + \beta_3\text{CT} + \beta_4\text{AGE} + \beta_5\text{SEX} + \beta_6\text{HT} \times \text{HS} + \beta_7\text{HT} \times \text{CT} + \beta_8\text{HT} \times \text{AGE} + \beta_9\text{HT} \times \text{SEX} + \beta_{10}\text{HS} \times \text{CT} + \beta_{11}\text{HS} \times \text{AGE} + \beta_{12}\text{HS} \times \text{SEX} + \beta_{13}\text{CT} \times \text{AGE} + \beta_{14}\text{CT} \times \text{SEX} + \beta_{15}\text{AGE} \times \text{SEX}.$
3. H_0: $\beta_6 = \beta_7 = \ldots = \beta_{15} = 0$, i.e., the coefficients of all product terms are zero.

   Likelihood ratio statistic: LR $= -2 \ln \hat{L}_1 - (-2 \ln \hat{L}_2)$, where $\hat{L}_1$ is the maximized likelihood for the reduced model (i.e., Exercise 1 model) and $\hat{L}_2$ is the maximized likelihood for the full model (i.e., Exercise 2 model).

   Distribution of LR statistic: chi square with 10 degrees of freedom.
4. H_0: $\beta_1 = 0$, where β_1 is the coefficient of HT in the no interaction model; alternatively, this null hypothesis can be stated as H_0: OR $= 1$, where OR denotes the odds ratio for the effect of HT adjusted for the other four variables in the no interaction model.

   Likelihood ratio statistic: LR $= -2 \ln \hat{L}_0 - (-2 \ln \hat{L}_1)$, where $\hat{L}_0$ is the maximized likelihood for the reduced model (i.e., Exercise 1 model less the HT term and its corresponding coefficient) and $\hat{L}_1$ is the maximized likelihood for the full model (i.e., Exercise 1 model).

   Distribution of LR statistic: approximately chi square with one degree of freedom.
5. The null hypothesis for the Wald test is the same as that given for the likelihood ratio test in Exercise 4. H_0: $\beta_1 = 0$ or, equivalently, H_0: OR $= 1$, where OR denotes the odds ratio for the effect of HT adjusted for the other four variables in the no interaction model.

   Wald test statistic: $Z=\hat{\beta}_1/s_{\hat{\beta}_1}$, where β_1 is the coefficient of HT in the no interaction model.

   Distribution of Wald statistic: approximately normal (0, 1) under H_0; alternatively, the square of the Wald statistic, i.e., Z^2, is approximately chi square with one degree of freedom.
6. The sample size for this study is 12,742, which is very large; consequently, the Wald and LR test statistics should be approximately the same.
7. The odds ratio of interest is given by e^{β_1}, where β_1 is the coefficient of HT in the no interaction model; a 95% confidence interval for this odds ratio is given by the following formula:

$$\exp\left[\hat{\beta}_1 \pm 1.96\sqrt{\widehat{\text{var}}\left(\hat{\beta}_1\right)}\right],$$

where $\widehat{\text{var}}(\hat{\beta}_1)$ is obtained from the variance–covariance matrix or, alternatively, by squaring the value of the standard error for $\hat{\beta}_1$ provided by the computer in the listing of variables and their estimated coefficients and standard errors.

8. H_0: $\beta_{\text{CAT}} = 0$ in the no interaction model (Model I), or alternatively, H_0: OR $= 1$, where OR denotes the odds ratio for the effect of CAT on CHD status, adjusted for the five other variables in Model I.

   Test statistic: Wald statistic $Z = \hat{\beta}_{\text{CAT}}/s_{\hat{\beta}_{\text{CAT}}}$, which is approximately normal (0, 1) under H_0, or alternatively, Z^2 is approximately chi square with one degree of freedom under H_0.

   Test computation: $Z = 0.5978/0.3520 = 1.70$; alternatively, $Z^2 = 2.88$; the one-tailed P-value is $0.0894/2 = 0.0447$, which is significant at the 5% level.

9. The point estimate of the odds ratio for the effect of CAT on CHD adjusted for the other variables in Model I is given by $e^{0.5978} = 1.82$. The 95% interval estimate for the above odds ratio is given by

$$\begin{aligned}\exp\left[\hat{\beta}_{\text{CAT}} \pm 1.96\sqrt{\widehat{\text{var}}\left(\hat{\beta}_{\text{CAT}}\right)}\right] &= (0.5978 \pm 1.96 \times 0.3520)\\ &= \exp(0.5978 \pm 0.6899)\\ &= \left(e^{-0.0921}, e^{1.2876}\right)\\ &= (0.91,\ 3.62).\end{aligned}$$

10. The null hypothesis for the likelihood ratio test for the effect of CC: H_0: $\beta_{\text{CC}} = 0$ where β_{CC} is the coefficient of CC in model II.

    Likelihood ratio statistic: LR = $-2 \ln \hat{L}_{\text{I}} - (-2 \ln \hat{L}_{\text{II}})$ where $\hat{L}_{\text{I}}$ and $\hat{L}_{\text{II}}$ are the maximized likelihood functions for Models I and II, respectively. This statistic has approximately a chi-square distribution with one degree of freedom under the null hypothesis.

    Test computation: LR $= 400.4 - 357.0 = 43.4$. The P-value is 0.0000 to four decimal places. Because P is very small, the null hypothesis is rejected and it is concluded that there is a significant effect of the CC variable, i.e., there is significant interaction of CHL with CAT.

11. The null hypothesis for the Wald test for the effect of CC is the same as that for the likelihood ratio test: H_0: $\beta_{\text{CC}} = 0$, where β_{CC} is the coefficient of CC in model II.

    Wald statistic: $Z = \hat{\beta}_{\text{CC}}/s_{\hat{\beta}_{\text{CC}}}$, which is approximately normal (0, 1) under H_0, or alternatively, Z^2 is approximately chi square with one degree of freedom under H_0.

Test computation: $Z = 0.0683/0.0143 = 4.77$; alternatively, $Z^2 = 22.75$; the two-tailed P-value is 0.0000, which is very significant.

The LR statistic is 43.4, which is almost twice as large as the square of the Wald statistic; however, both statistics are very significant, resulting in the same conclusion of rejecting the null hypothesis.

Model II is more appropriate than Model I because the test for interaction is significant.

12. The formula for the estimated odds ratio is given by

$$\begin{aligned}\widehat{\text{OR}}_{\text{adj}} &= \exp\left(\hat{\beta}_{\text{CAT}} + \hat{\delta}_{\text{CC}}\,\text{CHL}\right)\\ &= \exp(-14.089 + 0.0683\,\text{CHL}),\end{aligned}$$

where the coefficients come from Model II and the confounding effects of AGE, CHL, ECG, SMK, and HPT are adjusted.

13. Using the adjusted odds ratio formula given in Exercise 12, the estimated odds ratio values for CHL equal to 220 and 240 are:

$$\begin{aligned}\text{CHL} = 220\text{: } &\exp[-14.0809 + 0.0683(220)]\\ &= \exp(0.9451) = 2.57\end{aligned}$$

$$\begin{aligned}\text{CHL} = 240\text{: } &\exp[-14.0809 + 0.0683(240)]\\ &= \exp(2.3111) = 10.09\end{aligned}$$

14. Formula for the 95% confidence interval for the adjusted odds ratio when CHL = 220:

$$\exp\left[\hat{l} \pm 1.96\sqrt{\widehat{\text{var}}(\hat{l})}\right], \text{ where } \hat{l} = \hat{\beta}_{\text{CAT}} + \hat{\delta}_{\text{CC}}(220)$$

$$\begin{aligned}\text{and} \quad \widehat{\text{var}}(\hat{l}) &= \widehat{\text{var}}(\hat{\beta}_{\text{CAT}}) + (220)^2\,\widehat{\text{var}}(\hat{\delta}_{\text{CC}})\\ &\quad + 2(220)\,\widehat{\text{cov}}(\hat{\beta}_{\text{CAT}}, \hat{\delta}_{\text{CC}}),\end{aligned}$$

where $\widehat{\text{var}}(\hat{\beta}_{\text{CAT}}), \widehat{\text{var}}(\hat{\delta}_{\text{CC}})$, and $\widehat{\text{cov}}(\hat{\beta}_{\text{CAT}}, \hat{\delta}_{\text{CC}})$ are obtained from the printout of the variance– covariance matrix.

6 Modeling Strategy Guidelines

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_6, © Springer Science+Business Media, LLC 2010

Introduction

We begin this chapter by giving the rationale for having a strategy to determine a "best" model. Focus is on a logistic model containing a single dichotomous exposure variable that adjusts for potential confounding and potential interaction effects of covariates considered for control. A strategy is recommended, which has three stages: (1) variable specification, (2) interaction assessment, and (3) confounding assessment followed by consideration of precision. Causal diagrams are introduced as a component of the variable specification stage. The initial model must be "hierarchically well-formulated", a term to be defined and illustrated. Given an initial model, we recommend a strategy involving a "hierarchical backward elimination procedure" for removing variables. In carrying out this strategy, statistical testing is allowed for assessing interaction terms but is not allowed for assessing confounding. Further description of interaction and confounding assessment is given in the next chapter (Chap. 7).

Abbreviated Outline

The outline below gives the user a preview of the material in this chapter. A detailed outline for review purposes follows the presentation.

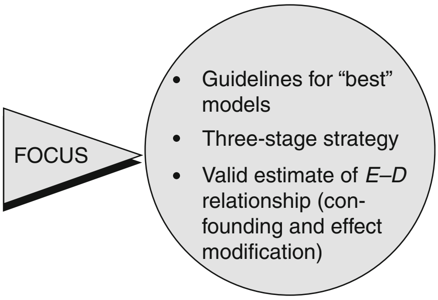
Objectives

Upon completion of this chapter, the learner should be able to:

1. State and recognize the three stages of the recommended modeling strategy.
2. Describe and/or illustrate a causal diagram that indicates confounding.
3. Define and recognize a hierarchically well-formulated logistic model.
4. State, recognize, and apply the recommended strategy for choosing potential confounders in one's model.
5. State, recognize, and apply the recommended strategy for choosing potential effect modifiers in one's model.
6. State and recognize the rationale for a hierarchically well-formulated model.
7. State and apply the hierarchical backward elimination strategy.
8. State and apply the hierarchy principle for retaining variables.
9. State whether or not significance testing is allowed for the assessment of interaction and/or confounding.

Presentation

I. Overview



- Guidelines for "best" models
- Three-stage strategy
- Valid estimate of *E–D* relationship (confounding and effect modification)

This presentation gives guidelines for determining the "best" model when carrying out mathematical modeling using logistic regression. We focus on a strategy involving three stages. The goal of this strategy is to obtain a valid estimate of an exposure–disease relationship that accounts for confounding and effect modification.

II. Rationale for a Modeling Strategy

We begin by explaining the rationale for a modeling strategy.

Minimum information in most study reports,
e.g., little explanation about strategy

Most epidemiologic research studies in the literature, regardless of the exposure–disease question of interest, provide a minimum of information about modeling methods used in the data analysis. Typically, only the final results from modeling are reported, with little accompanying explanation about the strategy used in obtaining such results.

Information often *not* provided:

- How variables are chosen
- How variables are selected
- How effect modifiers are assessed
- How confounders are assessed

For example, information is often *not* provided as to how variables are chosen for the initial model, how variables are selected for the final model, and how effect modifiers and confounders are assessed for their role in the final model.

Guidelines are needed for the following:

- To assess validity of results
- To help researchers know what information to provide
- To encourage consistency in strategy
- For a variety of modeling procedures

Without meaningful information about the modeling strategy used, it is difficult to assess the validity of the results provided. Thus, there is a need for guidelines regarding modeling strategy to help researchers know what information to provide.

In practice, most modeling strategies are ad hoc; in other words, researchers often make up a strategy as they go along in their analysis. The general guidelines that we recommend here encourage more consistency in the strategy used by different researchers.

Guidelines applicable to:

Logistic regression
Multiple linear regression
Cox PH regression

Modeling strategy guidelines are also important for modeling procedures other than logistic regression. In particular, classical multiple linear regression and Cox proportional hazards regression, although having differing model forms, all have in common with logistic regression the goal of describing exposure–disease relationships when used in epidemiologic research. The strategy offered here, although described in the context of logistic regression, is applicable to a variety of modeling procedures.

Two modeling goals:

(1) To obtain a valid E–D estimate
(2) To obtain a good predictive model

(different strategies for different goals)

There are typically two goals of mathematical modeling: One is to obtain a valid estimate of an exposure–disease relationship and the other is to obtain a good predictive model. Depending on which of these is the primary goal of the researcher, different strategies for obtaining the "best" model are required.

Prediction goal:

Use computer algorithms

When the goal is "prediction", it may be more appropriate to use computer algorithms, such as backward elimination or all possible regressions, which are built into computer packages for different models. [See Kleinbaum et al. (2008)]

Validity goal:

- Our focus
- For etiologic research
- Standard computer algorithms not appropriate

Our focus in this presentation is on the goal of obtaining a valid measure of effect. This goal is characteristic of most etiologic research in epidemiology. For this goal, standard computer algorithms do not apply because the roles that variables – such as confounders and effect modifiers – play in the model must be given special attention.

III. Overview of Recommended Strategy

Three stages:

(1) Variable specification
(2) Interaction assessment
(3) Confounding assessment followed by precision

The modeling strategy we recommend involves three stages: (1) *variable specification*, (2) *interaction assessment*, and (3) *confounding assessment followed by consideration of precision*. We have listed these stages in the order that they should be addressed.

Variable specification:

- Restricts attention to clinically or biologically meaningful variables
- Provides largest possible initial model

Variable specification is addressed first because this step allows the investigator to use the research literature to restrict attention to clinically or biologically meaningful independent variables of interest. These variables can then be defined in the model to provide the largest possible meaningful model to be initially considered.

Interaction prior to confounding:

- If strong interaction, then confounding irrelevant

Interaction assessment is carried out next, prior to the assessment of confounding. The reason for this ordering is that if there is strong evidence of interaction involving certain variables, then the assessment of confounding involving these variables becomes irrelevant.

EXAMPLE

Suppose *gender* is effect modifier for *E–D* relationship:

$$\underbrace{\widehat{\text{OR}}\text{ males} = 5.4,\ \widehat{\text{OR}}\text{ females} = 1.2}_{\text{interaction}}$$

Overall average = 3.5
not appropriate

Misleading because of separate effects for males and females

For example, suppose we are assessing the effect of an exposure variable *E* on some disease *D*, and we find strong evidence that *gender* is an effect modifier of the *E–D* relationship. In particular, suppose that the odds ratio for the effect of *E* on *D* is 5.4 for males but only 1.2 for females. In other words, the data indicate that the *E–D* relationship is different for males than for females, that is, there is interaction due to gender.

For this situation, it would *not* be appropriate to combine the two odds ratio estimates for males and females into a single overall adjusted estimate, say 3.5, that represents an "average" of the male and female odds ratios. Such an overall "average" is used to control for the confounding effect of gender in the absence of interaction; however, if interaction is present, the use of a single adjusted estimate is a misleading statistic because it masks the finding of a separate effect for males and females.

Assess interaction before confounding

Thus, we recommend that if one wishes to assess interaction and also consider confounding, then the assessment of interaction comes first.

Interaction may not be of interest:

- Skip interaction stage
- Proceed directly to confounding

However, the circumstances of the study may indicate that the assessment of interaction is not of interest or is biologically unimportant. In such situations, the interaction stage of the strategy can then be skipped, and one proceeds directly to the assessment of confounding.

EXAMPLE

Study goal: single overall estimate. Then interaction not appropriate

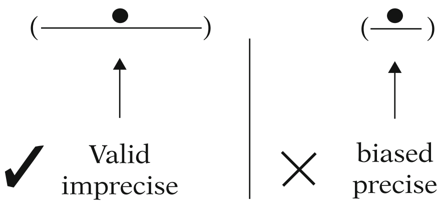
For example, the goal of a study may be to obtain a *single* overall estimate of the effect of an exposure adjusted for several factors, regardless of whether or not there is interaction involving these factors. In such a case, then, interaction assessment is not appropriate.

If interaction present:

- Do not assess confounding for effect modifiers
- Assessing confounding for other variables difficult and subjective

On the other hand, if interaction assessment is considered worthwhile, and, moreover, if significant interaction is found, then this precludes assessing confounding for those variables identified as effect modifiers. Also, as we will describe in more detail later, assessing confounding for variables other than effect modifiers can be quite difficult and, in particular, extremely subjective, when interaction is present.

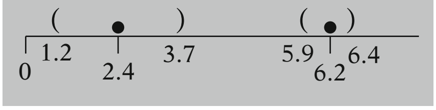
Confounding followed by precision:



The final stage of our strategy calls for the assessment of confounding followed by consideration of *precision*. This means that it is more important to get a valid point estimate of the *E–D* relationship that controls for confounding than to get a narrow confidence interval around a biased estimate that does not control for confounding.

EXAMPLE

Control Variables	$a\hat{OR}$	95% CI
✓AGE, RACE, SEX	2.4	(1.2, 3.7)
	↑	↑
	VALID	wide
AGE	6.2	(5.9, 6.4)
	↑	↑
	BIASED	narrow



For example, suppose controlling for *AGE*, *RACE*, and *SEX* simultaneously gave an adjusted odds ratio estimate of 2.4 with a 95% confidence interval ranging between 1.2 and 3.7, whereas controlling for *AGE alone* gave an odds ratio of 6.2 with a 95% confidence interval ranging between 5.9 and 6.4.

Then, assuming that *AGE*, *RACE*, and *SEX* are *considered important risk factors* for the disease of interest, we would prefer to use the odds ratio of 2.4 over the odds ratio of 6.2. This is because the 2.4 value results from controlling for all the relevant variables and, thus, gives us a more valid answer than the value of 6.2, which controls for only one of the variables.

Thus, even though there is a much narrower confidence interval around the 6.2 estimate than around the 2.4, the gain in precision from using 6.2 does not offset the bias in this estimate when compared to the more valid 2.4 value.

VALIDITY BEFORE PRECISION
↓ ↓
right answer precise answer

In essence, then, *validity takes precedence over precision, so that it is more important to get the right answer than a precise answer*. Thus, in the third stage of our strategy, we seek an estimate that controls for confounding and is, over and above this, as precise as possible.

Confounding : *no statistical testing*

↓

Validity — systematic error

(Statistical testing — random error)

When later describing this last stage in more detail we will emphasize that *the assessment of confounding is carried out without using statistical testing*. This follows from general epidemiologic principles in that confounding is a validity issue that addresses systematic rather than random error. Statistical testing is appropriate for considering random error rather than systematic error.

Confounding in logistic regression — a validity issue

Computer algorithms no good (involve statistical testing)

Our suggestions for assessing confounding using logistic regression are consistent with the principle that confounding is a validity issue. Standard computer algorithms for variable selection, such as forward inclusion or backward elimination procedures, are not appropriate for assessing confounding because they involve statistical testing.

Statistical issues beyond the scope of this presentation:

- Multicollinearity
- Multiple testing
- Influential observations

Before concluding this overview section, we point out a few statistical issues needing attention but which are beyond the scope of this presentation. These issues are *multicollinearity*, *multiple testing*, and *influential observations*.

Multicollinearity:

- Independent variables approximately determined by other independent variables
- Regression coefficients unreliable

Multicollinearity occurs when one or more of the independent variables in the model can be approximately determined by some of the other independent variables. When there is multicollinearity, the estimated regression coefficients of the fitted model can be highly unreliable. Consequently, any modeling strategy must check for possible multicollinearity at various steps in the variable selection process.

Multiple testing:

- The more tests, the more likely significant findings, even if no real effects
- Variable selection procedures may yield an incorrect model because of multiple testing

Multiple testing occurs from the many tests of significance that are typically carried out when selecting or eliminating variables in one's model. The problem with doing several tests on the same data set is that the more tests one does, the more likely one can obtain statistically significant results even if there are no real associations in the data. Thus, the process of variable selection may yield an incorrect model because of the number of tests carried out. Unfortunately, there is no foolproof method for adjusting for multiple testing, even though there are a few rough approaches available.

Influential observations:

- Individual data may influence regression coefficients, e.g., outlier
- Coefficients may change if outlier is dropped from analysis

Influential observations refer to data on individuals that may have a large influence on the estimated regression coefficients. For example, an outlier in one or more of the independent variables may greatly affect one's results. If a person with an outlier is dropped from the data, the estimated regression coefficients may greatly change from the coefficients obtained when that person is retained in the data. Methods for assessing the possibility of influential observations should be considered when determining a best model.

IV. Variable Specification Stage

- Define clinically or biologically meaningful independent variables
- Provide initial model

At the variable specification stage, clinically or biologically meaningful independent variables are defined in the model to provide the largest model to be initially considered.

Specify D, E, C_1, C_2, $\ldots$, C_p based on:

- Study goals
- Literature review
- Theory

We begin by specifying the D and E variables of interest together with the set of risk factors C_1 through C_p to be considered for control. These variables are defined and measured by the investigator based on the goals of one's study and a review of the literature and/or biological theory relating to the study.

Specify Vs based on:

- Prior research or theory
- Possible statistical problems

Next, we must specify the Vs, which are functions of the Cs that go into the model as potential confounders. Generally, we recommend that the choice of Vs be based primarily on prior research or theory, with some consideration of possible statistical problems like multicollinearity that might result from certain choices.

EXAMPLE

Cs: AGE, RACE, SEX
Vs:
Choice 1: AGE, RACE, SEX

Choice 2: AGE, RACE, SEX, AGE^2,
AGE × RACE, RACE × SEX,
AGE × SEX

For example, if the Cs are AGE, RACE, and SEX, one choice for the Vs is the Cs themselves. Another choice includes AGE, RACE, and SEX plus more complicated functions such as AGE^2, AGE × RACE, RACE × SEX, and AGE × SEX.

We would recommend any of the latter four variables only if prior research or theory supported their inclusion in the model. Moreover, even if biologically relevant, such variables may be omitted from consideration to avoid a possible collinearity problem.

✓ Simplest choice for Vs:

The Cs themselves (or a subset of Cs)

The simplest choice for the Vs is the Cs themselves. If the number of Cs is very large, it may even be appropriate to consider a smaller subset of the Cs considered to be most relevant and interpretable based on prior knowledge.

Specify Ws: (in model as $E \times W$):

Restrict Ws to be Vs themselves or products of two Vs

(i.e., in model as $E \times V$ and $E \times V_i \times V_j$)

Once the Vs are chosen, the next step is to determine the Ws. These are the effect modifiers that go into the model as product terms with E, that is, these variables are of the form E times W.

We recommend that the choice of Ws be restricted either to the Vs themselves or to product terms involving two Vs. Correspondingly, the product terms in the model are recommended to be of the form E times V and E times V_i times V_j, where V_i and V_j are two distinct Vs.

Most situations:

Specify Vs and Ws as Cs or subset of Cs

For most situations, we recommend that both the Vs and the Ws be the Cs themselves, or even a subset of the Cs.

EXAMPLE

$C_1, C_2, C_3,$ = AGE, RACE, SEX

$V_1, V_2, V_3,$ = AGE, RACE, SEX

Ws = subset of AGE, RACE, SEX

As an example, if the Cs are AGE, RACE, and SEX, then a simple choice would have the Vs be AGE, RACE, and SEX and the Ws be a subset of AGE, RACE, and SEX thought to be biologically meaningful as effect modifiers.

Rationale for Ws (common sense):

Product terms more complicated than EV_iV_j are as follows:

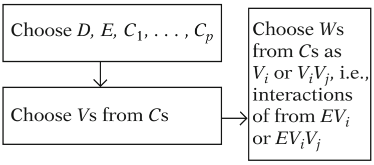
- Difficult to interpret
- Typically cause collinearity

✓ Simplest choice: use EV_i terms only

The *rationale* for our recommendation about the Ws is based on the following commonsense considerations:

- Product terms more complicated than EV_iV_j are usually *difficult to interpret* even if found significant; in fact, even terms of the form EV_iV_j are often uninterpretable.
- Product terms more complicated than EV_iV_j typically will cause *collinearity* problems; this is also likely for EV_iV_j terms, so the simplest way to reduce the potential for multicollinearity is to use EV_i terms only.

Variable Specification Summary Flow Diagram


Choose D, E, $C_1, \ldots, C_p$
Choose Vs from Cs
Choose Ws from Cs as V_i or V_iV_j, i.e., interactions of from EV_i or EV_iV_j

In summary, at the variable specification stage, the investigator defines the largest possible model initially to be considered. The flow diagram at the left shows first the choice of D, E, and the Cs, then the choice of the Vs from the Cs and, finally, the choice of the Ws in terms of the Cs.

V. Causal Diagrams

Approach for variable selection:

- Not just quantitative
- Consider causal structure
- Depends on the goal

The decision of specifying which variables are potential confounders should not just be based on quantitative methods; we must also consider the possible causal relationships between the exposure, outcome, potential confounders, and other relevant variables. Moreover, we must be clear about the goal of our analysis.

Including covariate in model could lead to bias:

- If caused by exposure
- If caused by outcome

Including a variable in the model that is associated with the outcome could lead to bias of the exposure–disease relationship if the level of that variable *was caused* by the exposure and/or by the outcome.

Lung cancer causes an abnormal X-ray, i.e.,

Lung cancer $(D) \rightarrow$ Chest X-ray (C)

We claim:

E, D association controlling for C is biased

Finding an abnormal X-ray could be a consequence of lung cancer (we have indicated this graphically by the one-sided arrow on the left – a simple example of a *causal diagram*). If we were interested in estimating the causal association between cigarette smoking and lung cancer (as opposed to developing our best predictive model of lung cancer), it would bias our results to include chest X-ray status as a covariate.

Model:

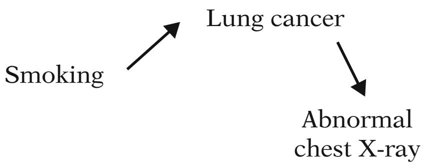
$$\text{logit } P(D = 1|\mathbf{X}) = \beta_0 + \beta_1\text{SMOKE} + \beta_2\text{XRY}$$

More specifically, consider a logistic model with lung cancer as the outcome and smoking status and chest X-ray status as covariates (model stated on the left).

Where D coded 1 for lung cancer
0 for no lung cancer
SMOKE coded 1 for smokers,
0 for nonsmokers
XRY coded 1 for abnormal X-ray
0 for normal X-ray

$\mathbf{exp}(\beta_1) = \text{OR}(\text{SMOKE} = 1 \text{ vs. } 0)$ holding X-ray status constant

Now consider the interpretation of the odds ratio for SMOKE derived from this model, **exp**(β_1); i.e., the odds of lung cancer among the smokers divided by the odds of lung cancer among the nonsmokers, *holding X-ray status constant* (i.e., adjusting for X-ray status).



Above causal diagram
$\Downarrow$
Bias if we condition on X-ray status

The causal diagram at the left describes the likely causal pathway that involves the three variables smoking, lung cancer, and abnormal chest X-ray.

We can use this diagram to explain why any association between smoking and lung cancer is weakened (and therefore biased) if we control for X-ray status. In particular, a consequence of the causal effect of smoking on lung cancer is to increase the likelihood of an abnormal X-ray.

Explanation:
Among smokers with abnormal chest X-ray
- High odds of lung cancer

Among nonsmokers with abnormal X-ray
- High odds of lung cancer

Among those with abnormal chest X-ray
- Odds ratio (smoking vs. nonsmoking) closer to null than in general population (a bias)

Explaining the reason for this bias, we would expect a large proportion of smokers who have an abnormal chest X-ray to have lung cancer simply because an abnormal X-ray is a strong indicator of lung cancer. However, we would also expect a large proportion of nonsmokers who have an abnormal chest X-ray to have lung cancer. So among those who have an abnormal chest X-ray, the odds of lung cancer would not substantially differ comparing smokers to nonsmokers, even though the odds would differ greatly in the general population.

Depending on the underlying causal structure, adjustment may:
- Remove bias
- Lead to bias
- Neither of the above

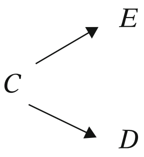
The point of the above example is that even though the adjusted odds ratio may be much different than the unadjusted odds ratio, adjustment may cause bias rather than remove bias. Whether adjustment causes bias or removes bias (or neither), depends on the underlying causal structure of the variables of interest.

Causal diagrams may help understanding of:
- Causation
- Association
- Bias

Causal diagrams provide a graphical perspective for understanding epidemiologic concepts involving causation, association, and bias. In this section, we highlight the key concepts. A more detailed description of causal diagrams can be found elsewhere (Rothman et al., 2008).

Causal Diagram for Confounding

C is a common cause of *E* and *D*

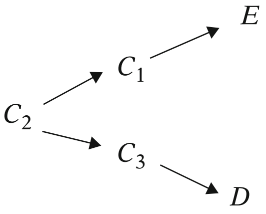


Noncausal *E*–*D* association
C confounds the *E*–*D* relationship

The path *E*–*C*–*D* is a *backdoor path* from *E* to *D*

Confounding of the exposure–disease association is rooted in a common cause (*C*) of the exposure (*E*) and disease (*D*), leading to a spurious *E*–*D* association.

The diagram on the left illustrates that *E* does not cause *D*, yet there is a noncausal pathway between *E* and *D* through *C*. Such a noncausal pathway between two variables of interest is called a "backdoor path". The noncausal path from *E* to *D* goes through *C* and is denoted as *E*–*C*–*D*.

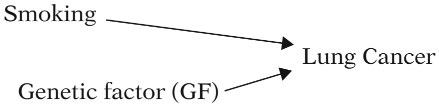


E–C_1–C_2–C_3–D is a backdoor path

Can control for either C_1 or C_2 or C_3

The next diagram (on the left) is somewhat more complicated. C_2 is a common cause of *E* (through C_1) and *D* (through C_3). A noncausal backdoor path E–C_1–C_2–C_3–D will lead to a spurious association between *E* and *D* if not adjusted. Although C_2 is the common cause, you can control for (condition on) either C_1 or C_2 or C_3. A confounder need not be the common cause; it just needs to be on the path to or from the common cause.

Lung cancer a common effect



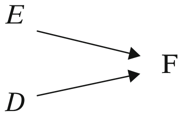
The next type of causal structure we examine is one that contains a common effect from two or more causes. Consider two independent risk factors for lung cancer: smoking and some genetic factor (GF). As shown on the left, lung cancer is a common effect of these risk factors.

In general population:
No association between GF and smoke
Among lung cancer patients:

- Smokers may get lung cancer because they smoke
- Smoking is not a reason that a nonsmoker gets lung cancer (omitting secondhand smoke as a reason)
- So nonsmokers more likely to have genetic factor than smokers (smoking associated with GF among lung cancer patients)

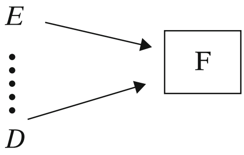
Suppose there is no association between smoking and the genetic factor in the general population. Nevertheless, among lung cancer patients, there likely is an association between smoking and the genetic factor. Nonsmokers who get lung cancer get lung cancer for some reason. Since smoking is not the reason they got lung cancer, nonsmokers may be more likely to have the genetic factor as the reason compared to smokers who get lung cancer.

F is a common effect of *E* and *D*:

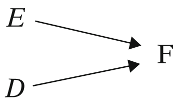


This spurious association produced by conditioning on a common effect can be expressed with causal diagrams. Let F be a common effect of the exposure (*E*) and disease (*D*) with exposure unrelated to disease.

Conditioning on F creates a spurious association between *E* and *D*



The second causal diagram, with the box around F (the common effect), indicates conditioning, or adjusting, on F. The dotted lines between *E* and *D* without a causal arrow indicate that a spurious association between *E* and *D* was produced because of the conditioning on F (i.e., within strata of F).

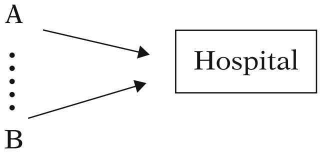


Backdoor path *E–F–D* is blocked by common effect. No spurious association unless we condition on F.

If we do not condition on a common effect we may still wonder if there is a spurious association between *E* and *D* because of the backdoor path E–*F*–*D*. However, a backdoor path through a common effect will *not* create a spurious association, unless we condition on that common effect.

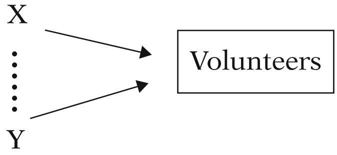
Berkson's bias:

Selecting only hospital patients could lead to bias of A–B association.



Joseph Berkson illustrated this bias in studies in which selected subjects were hospitalized patients (Berkson, 1946). If condition A and condition B can lead to hospitalization, then selecting only hospitalized patients can yield a biased estimated association between A and B.

Selecting volunteers could lead to bias of X–Y association.



Similarly, if factors X and Y influenced volunteerism, then restricting the study population to volunteers could lead to a selection bias of the X–Y association.

Conditioning on a common cause can

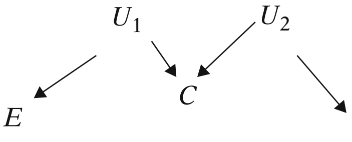
- Remove bias

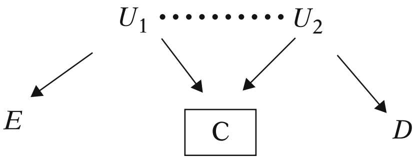
Conditioning on a common effect can

- Induce bias

We have seen that conditioning on a common cause (a confounder) can remove bias and conditioning on a common effect can induce bias.

U_1 and U_2 are unmeasured
Should we control for C?





By controlling for C we create an unblocked path from E to D: E–U_1–U_2–D

Do not control for C

For a more complicated example, consider the causal diagram on the left. Suppose U_1 and U_2 are unmeasured factors, with U_1 being a common cause of E and C, and with U_2 being a common cause of D and C. If we are interested in estimating an unbiased measure of effect between E and D, should we control for C?

U_1 is a cause of E, and U_2 is a cause of D but there is no common cause of E and D, thus there is no confounding. However, if we condition on C, a common effect of U_1 and U_2, then we create a link between U_1 and U_2 (i.e., a spurious association) and an unblocked backdoor path from E to D leading to a spurious association between E and D. The backdoor path is E–U_1–U_2–D. Since U_1 and U_2 are unmeasured we cannot adjust for either of these variables and block that backdoor path. Therefore, we should not control for C.

Is it too much to expect that we correctly and completely specify the underlying causal structure?
Answer : Yes

Correctly specifying the causal structure of all the relevant variables for assessing the E–D relationship is close to impossible. However, this does not mean that we should not think about the underlying causal structure.

Do we run the risk of inducing bias if we do not consider the causal structure at all?
Answer : Yes

We should certainly be aware that decisions to include or not include covariates in the model may induce or remove bias depending on the causal relationships. In particular, we should be aware that conditioning on a common effect can induce bias.

Analytic goal:

- Estimate E–D relationship
  ⇒ Concern about causal structure confounding, interaction
- Predict the outcome
  ⇒ Causal structure of less concern

Central to this discussion and to all our discussion on model strategy is that our goal is to obtain a valid estimate of an exposure–disease relationship. If our goal was to obtain the best predictive model, we would not be so concerned about the causal structure, confounding, or interaction.

VI. Other Considerations for Variable Specification

There are other issues that need to be considered at the variable specification stage. We briefly discuss them in this section.

Data quality:

Measurement error, misclassification?

Correct or remove missing data?

First, we should consider the quality of the data: Does the variable contain the information we want? Is there an unacceptable level of measurement error or misclassification? What is the number of missing observations? If an observation is missing for *any* covariate in a model, typically computer programs "throw out" that observation when running that model.

(Qualitative) Collinearity:

Are covariates supplying qualitatively redundant info?

We should also consider whether there is collinearity between covariates. In this context, we are not considering collinearity as a model diagnostic as we describe quantitatively in Chap. 8. Rather, here we are considering whether two covariates are qualitatively redundant.

Example:

Including both employment status and personal income in model.

Controlling for same underlying factor?
(leads to model instability)

Controlling for meaningfully different factors?
(needed for proper control)

For example, suppose we include two variables in a model to control for both employment status and personal income. If these two variables control the same underlying factor, then including them both in the same model could lead to model instability. On the other hand, if you believe that employment status and personal income are meaningfully different, then including them both may be important for proper control.

Sample size?
If large ⇒ can "tease out" effects of similar covariates

A consideration of whether a model can include similar, but not identical covariates, is the sample size of the data. A large dataset can better support the "teasing out" of subtle effects compared with a dataset with a relatively small number of observations.

Philosohical issue: *complexity vs. simplicity*

- Complexity: If in doubt, include the variable. Better safe than sorry.
- Simplicity – If in doubt, keep it out. It is a virtue to be simple.

Another consideration is philosophical. Some prefer simplicity – if in doubt, leave the variable out. Others say – if in doubt, include the variable as it is better to be safe than sorry. Albert Einstein is attributed to have said "keep everything as simple as possible, but not simpler."

Get to know your data!

Perform thorough descriptive analyses before modeling.

- Useful for finding data errors
- Gain insight about your data

It is important to do thorough descriptive analyses before modeling. Get to know your data! It is possible to run many models and not know that you have only two smokers in your dataset. Also descriptive analyses are useful for finding errors. An individual's age may be incorrectly recorded at 699 rather than 69 and you may never know that from reading model output.

Descriptive analyses include the following:

- Frequency tables
- Summary statistics
- Correlations
- Scatter plots
- Histograms

Descriptive analyses include obtaining frequency tables for categorical variables, univariate summary statistics (means, variance, quartiles, max, min, etc.) for continuous variable, bivariate cross tables, bivariate correlations, scatter plots, and histograms. Descriptive analyses can be performed both before and after the variable specification stage. Often more insight is gained from a descriptive analysis than from modeling.

VII. Hierarchically Well-Formulated Models

Initial model structure: HWF

When choosing the V and W variables to be included in the initial model, the investigator must ensure that the model has a certain structure to avoid possibly misleading results. This structure is called a *hierarchically well-formulated model*, abbreviated as HWF, which we define and illustrate in this section.

Model contains *all lower-order components*

A hierarchically well-formulated model is a model satisfying the following characteristic: Given any variable in the model, all lower-order components of the variable must also be contained in the model.

EXAMPLE

Not HWF model:

$$\text{logit P}\left(\mathbf{X}\right) = \alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 + \delta_1 EV_1 + \delta_2 EV_2 + \delta_3 EV_1 V_2$$

Components of EV_1V_2:
E, V_1, V_2, EV_1, EV_2, $\underset{\uparrow \text{ not in model}}{V_1V_2}$

To understand this definition, let us look at an example of a model that is *not* hierarchically well formulated. Consider the model given in logit form as logit P($\mathbf{X}$) equals α plus βE plus $\gamma_1 V_1$ plus $\gamma_2 V_2$ plus the product terms $\delta_1 EV_1$ plus $\delta_2 EV_2$ plus $\delta_3 EV_1V_2$.

For this model, let us focus on the three-factor product term EV_1V_2. This term has the following lower-order components: E, V_1, V_2, EV_1, EV_2, and V_1V_2. Note that the last component V_1V_2 is not contained in the model. Thus, the model is not hierarchically well formulated.

EXAMPLE

HWF model:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 + \delta_1 E V_1 + \delta_2 E V_2$$

Components of EV_1:
E, V_1 both in model

In contrast, the model given by logit P(**X**) equals α plus βE plus $\gamma_1 V_1$ plus $\gamma_2 V_2$ plus the product terms $\delta_1 EV_1$ plus $\delta_2 EV_2$ is hierarchically well formulated because the lower-order components of each variable in the model are also in the model. For example, the components of EV_1 are E and V_1, both of which are contained in the model.

EXAMPLE

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1^2 + \gamma_2 V_2 + \delta_1 E V_1^2$$

HWF model?
Yes, if V_1^2 is biologically meaningful
components of EV_1^2 E and V_1^2
components of V_1^2: none

No, if V_1^2 is not meaningful separately from V_1:
The model does not contain
- V_1, component of V_1^2
- EV_1, component of EV_1^2

For illustrative purposes, let us consider one other model given by logit P(**X**) equals α plus βE plus $\gamma_1 V_1^2$ plus $\gamma_2 V_2$ plus the product term $\delta_1 EV_1^2$. Is this model hierarchically well formulated?

The answer here can be either *yes* or *no* depending on how the investigator wishes to treat the variable V_1^2 in the model. If V_1^2 is biologically meaningful in its own right without considering its component V_1, then the corresponding model is hierarchically well formulated because the variable EV_1^2 can be viewed as having only two components, namely, E and V_1^2, both of which are contained in the model. Also, if the variable V_1^2 is considered meaningful by itself, it can be viewed as having no lower order components. Consequently, all lower order components of each variable are contained in the model.

On the other hand, if the variable V_1^2 is not considered meaningful separately from its fundamental component V_1, then the model is not hierarchically well formulated. This is because, as given, the model does not contain V_1, which is a lower order component of V_1^2 and EV_1^2, and also does not contain the variable EV_1, which is a lower order component of EV_1^2.

Why require HWF model?

Answer:

HWF?	Tests for highest-order variables?
No	dependent on coding
Yes	independent of coding

Now that we have defined and illustrated an HWF model, we discuss why such a model structure is required. The reason is that if the model is not HWF, then tests about variables in the model – in particular, the highest-order terms – may give varying results depending on the coding of variables in the model. Such tests should be *independent of the coding* of the variables in the model, and they are if the model is hierarchically well formulated.

EXAMPLE

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 + \delta_1 E V_1 + \delta_2 E V_2 + \delta_3 E V_1 V_2$$

Not HWF model:
V_1V_2 missing

To illustrate this point, we return to the first example considered above, where the model is given by logit P(**X**) equals α plus βE plus $\gamma_1 V_1$ plus $\gamma_2 V_2$ plus the product terms $\delta_1 E V_1$ plus $\delta_2 E V_2$ plus $\delta_3 E V_1 V_2$. This model is not hierarchically well formulated because it is missing the term $V_1 V_2$. The highest-order term in this model is the three-factor product term EV_1V_2.

EXAMPLE (continued)

E dichotomous:

Then if *not* HWF model,
testing for EV_1V_2 may depend on whether E is coded as

$E = (0, 1)$, e.g., significant

or

$E = (-1, 1)$, e.g., not significant

or

other coding

Suppose that the exposure variable E in this model is a dichotomous variable. Then, because the model is not HWF, a test of hypothesis for the significance of the highest-order term, EV_1V_2, may give different results depending on whether E is coded as (0, 1) or (−1, 1) or any other coding scheme.

In particular, it is possible that a test for EV_1V_2 may be highly significant if E is coded as (0, 1), but be nonsignificant if E is coded as (−1, 1). Such a possibility should be avoided because the coding of a variable is simply a way to indicate categories of the variable and, therefore, should not have an effect on the results of data analysis.

EXAMPLE

HWF model:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 + \gamma_3 V_1 V_2 + \delta_1 E V_1 + \delta_2 E V_2 + \delta_3 E V_1 V_2$$

Testing for EV_1V_2 is *independent of coding* of E: (0, 1), (−1, 1), or other.

In contrast, suppose we consider the HWF model obtained by adding the V_1V_2 term to the previous model. For this model, a test for EV_1V_2 will give exactly the same result whether E is coded using (0, 1), (−1, 1), or any other coding. In other words, such a test is independent of the coding used.

HWF model. Tests for *lower* order terms depend on coding

We will shortly see that even if the model is hierarchically well formulated, then tests about lower order terms in the model may still depend on the coding.

EXAMPLE

HWF model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 + \gamma_3 V_1 V_2 + \delta_1 E V_1 + \delta_2 E V_2 + \delta_3 E V_1 V_2$$

EV_1V_2: *not dependent* on coding

EV_1 or EV_2: *dependent* on coding

For example, even though, in the HWF model being considered here, a test for EV_1V_2 is not dependent on the coding, a test for EV_1 or EV_2 – which are lower order terms – may still be dependent on the coding.

Require

- HWF model
- No test for lower order components of significant higher order terms

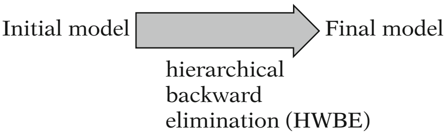
What this means is that in addition to requiring that the model be HWF, we also require that no tests be allowed for lower order components of terms like EV_1V_2 already found to be significant. We will return to this point later when we describe the "hierarchy principle" for retaining variables in the model.

VIII. The Hierarchical Backward Elimination Approach

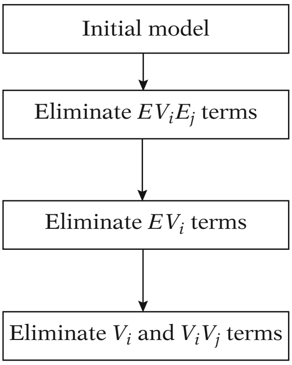
✓ Variable specification
✓ HWF model

Largest model considered = initial (starting) model

We have now completed our recommendations for variable specification as well as our requirement that the model be hierarchically well formulated. When we complete this stage, we have identified the largest possible model to be considered. This model is the initial or starting model from which we attempt to eliminate unnecessary variables.



The recommended process by which the initial model is reduced to a final model is called a *hierarchical backward elimination (HWBE) approach*. This approach is described by the flow diagram shown here.



In the flow diagram, we begin with the initial model determined from the variable specification stage.

If the initial model contains three-factor product terms of the form EV_iV_j, then we attempt to eliminate these terms first.

Following the three-factor product terms, we then eliminate unnecessary two-factor product terms of the form EV_i.

The last part of the strategy eliminates unnecessary V_i and V_iV_j terms.

EV_i and EV_iV_j (interactions): use statistical testing

As described in later sections, the EV_iV_j and EV_i product terms can be eliminated using appropriate statistical testing methods.

V_i and V_iV_j (confounders): do *not* use statistical testing

However, decisions about the V_i and V_iV_j terms, which are potential confounders, should not involve statistical testing.

Hierarchical

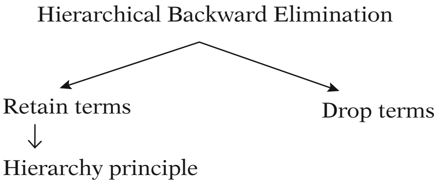
3 factors: EV_iV_j
↓
2 factors: EV_i
↓
2 factors: V_iV_j
↓
1 factors: V_i

Backward

Large starting model
↓
Smaller final model

The *strategy* described by this flow diagram is called *hierarchical backward* because we are working backward from our largest starting model to a smaller final and we are treating variables of different orders at different steps. That is, there is a hierarchy of variable types, with three-factor interaction terms considered first, followed by two-factor interaction terms, followed by two-factor, and then one-factor confounding terms.

IX. The Hierarchy Principle for Retaining Variables



As we go through the hierarchical backward elimination process, some terms are retained and some terms are dropped at each stage. For those terms that are retained at a given stage, there is a rule for identifying lower order components that must also be retained in any further models.

(Bishop, Fienberg, and Holland, 1975)

This rule is called the *hierarchy principle*. An analogous principle of the same name has been described by Bishop, Fienberg, and Holland (1975).

EXAMPLE

Initial model: $EV_i\ V_j$ terms

Suppose: EV_2V_5 significant

Hiearchy principle. all lower order components of EV_2V_5 retained

i.e., E, V_2, V_5, EV_2, EV_5, and V_2V_5 cannot be eliminated

Note. Initial model must contain V_2V_5 to be HWF

To illustrate the hierarchy principle, suppose the initial model contains three-factor products of the form EV_iV_j. Suppose, further, that the term EV_2V_5 is found to be significant during the stage that considers the elimination of unimportant EV_iV_j terms. Then, the hierarchy principle requires that all lower order components of the EV_2V_5 term must be retained in all further models considered in the analysis.

The lower order components of EV_2V_5 are the variables E, V_2, V_5, EV_2, EV_5, and V_2V_5. Because of the hierarchy principle, if the term EV_2V_5 is retained, then each of the above component terms cannot be eliminated from all further models considered in the backward elimination process. Note the initial model has to contain each of these terms, including V_2V_5, to ensure that the model is hierarchically well formulated.

Hierarchy Principle

If product variable retained, then *all* lower order components must be retained

In general, the hierarchy principle states that if a product variable is retained in the model, then all lower order components of that variable must be retained in the model.

EXAMPLE

EV_2 and EV_4 retained:
Then
E, V_2 and V_4 also retained

cannot be considered as nonconfounders

As another example, if the variables EV_2 and EV_4 are to be retained in the model, then the following lower order components must also be retained in all further models considered: E, V_2, and V_4. Thus, we are not allowed to consider dropping V_2 and V_4 as possible nonconfounders because these variables must stay in the model regardless.

Hiearchy principle rationale:

- Tests for lower order components depend on coding
- Tests should be independent of coding
- Therefore, no tests allowed for lower order components

The *rationale for the hierarchy principle* is similar to the rationale for requiring that the model be HWF. That is, tests about lower order components of variables retained in the model can give different conclusions depending on the coding of the variables tested. Such tests should be independent of the coding to be valid. Therefore, no such tests are appropriate for lower order components.

EXAMPLE

Suppose EV_2V_5 significant: then the test for EV_2 depends on coding of E, e.g., (0, 1) or (−1, 1)

For example, if the term EV_2V_5 is significant, then a test for the significance of EV_2 may give different results depending on whether E is coded as (0, 1) or (−1, 1).

HWF model:

Tests for *highest-order* terms *independent* of coding

but

tests for *lower order* terms *dependent* on coding

Note that if a model is HWF, then tests for the highest-order terms in the model are always independent of the coding of the variables in the model. However, tests for lower order components of higher order terms are still dependent on coding.

EXAMPLE

HWF: EV_iV_j highest-order terms
Then tests for
EV_iV_j *independent* of coding *but* tests for
EV_i or V_j *dependent* on coding

For example, if the highest-order terms in an HWF model are of the form EV_iV_j, then tests for all such terms are not dependent on the coding of any of the variables in the model. However, tests for terms of the form EV_i or V_i are dependent on the coding and, therefore, should not be carried out as long as the corresponding higher order terms remain in the model.

EXAMPLE

HWF: EV_i highest-order terms
Then tests for
EV_i *independent* of coding *but* tests for
V_i *dependent* on coding

If the highest-order terms of a HWF model are of the form EV_i, then tests for EV_i terms are independent of coding, but tests for V_i terms are dependent on coding of the *V*s and should not be carried out. Note that because the *V*s are potential confounders, tests for *V*s are not allowed anyhow.

- Ensures that the model is HWF

e.g., EV_iV_j is significant
⇒ retain lower order components or else model is not HWF

Note also, regarding the hierarchy principle, that any lower order component of a significant higher order term must remain in the model or else the model will no longer be HWF. Thus, to ensure that our model is HWF as we proceed through our strategy, we cannot eliminate lower order components unless we have eliminated corresponding higher order terms.

X. An Example

EXAMPLE

Cardiovascular Disease Study
9-year follow-up Evans County, GA
$n = 609$ white males

The variables:

$$\underbrace{\text{CAT, AGE, CHL, SMK, ECG, HPT}}_{\text{at start}}$$

CHD = outcome

We review the guidelines recommended to this point through an example. We consider a cardiovascular disease study involving the 9-year follow-up of persons from Evans County, Georgia. We focus on data involving 609 white males on which we have measured six variables at the start of the study. These are catecholamine level (CAT), AGE, cholesterol level (CHL), smoking status (SMK), electrocardiogram abnormality status (ECG), and hypertension status (HPT). The outcome variable is coronary heart disease status (CHD).

CAT: (0, 1) exposure

$$\left.\begin{array}{r}\text{AGE, CHL : continuous}\\ \text{SMK, ECG, HPT : (0, 1)}\end{array}\right\}\begin{array}{l}\text{control}\\ \text{variables}\end{array}$$

In this study, the exposure variable is CAT, which is 1 if high and 0 if low. The other five variables are control variables, so that these may be considered as confounders and/or effect modifiers. AGE and CHL are treated continuously, whereas SMK, ECG, and HPT, are (0, 1) variables.

$$E = \text{CAT} \xrightarrow{\qquad ?\qquad} D = \text{CHD}$$

controlling for

$$\underbrace{\text{AGE, CHL, SMK, ECG, HPT}}_{C\text{s}}$$

The question of interest is to describe the relationship between E (CAT) and D (CHD), controlling for the possible confounding and effect-modifying effects of AGE, CHL, SMK, ECG, and HPT. These latter five variables are the Cs that we have specified at the start of our modeling strategy.

Variable specification stage:
Vs: potential confounders in initial model

To follow our strategy for dealing with this data set, we now carry out variable specification in order to define the initial model to be considered. We begin by specifying the V variables, which represent the potential confounders in the initial model.

Here, Vs = Cs:

$V_1 = \text{AGE}, V_2 = \text{CHL}, V_3 = \text{SMK},$

$V_4 = \text{ECG}, V_5 = \text{HPT}$

In choosing the Vs, we follow our earlier recommendation to let the Vs be the same as the Cs. Thus, we will let $V_1 = \text{AGE}$, $V_2 = \text{CHL}$, $V_3 = \text{SMK}$, $V_4 = \text{ECG}$, and $V_5 = \text{HPT}$.

Other possible Vs:
$V_6 = \text{AGE} \times \text{CHL}$
$V_7 = \text{AGE} \times \text{SMK}$
$V_8 = \text{AGE}^2$
$V_9 = \text{CHL}^2$

We could have chosen other Vs in addition to the five Cs. For example, we could have considered Vs that are products of two Cs, such as V_6 equals AGE × CHL or V_7 equals AGE × SMK. We could also have considered Vs that are squared Cs, such as V_8 equals AGE^2 or V_9 equals CHL^2.

EXAMPLE (continued)

Restriction of Vs to Cs because:

- Large number of Cs
- Additional Vs difficult to interpret
- Additional Vs may lead to collinearity

Choice of Ws:
(go into model as EW)
Ws = Cs:

$$W_1 = \text{AGE}, W_2 = \text{CHL}, W_3 = \text{SMK},$$
$$W_4 = \text{ECG}, W_5 = \text{HPT}$$

Other possible Ws:
$W_6 = \text{AGE} \times \text{CHL}$

(If W_6 is in model, then
$V_6 = \text{AGE} \times \text{CHL}$ also in HWF model.)

Alternative choice of Ws:
Subset of Cs, e.g.,

AGE ⇒ CAT × AGE in model
ECG ⇒ CAT × ECG in model

Rationale for Ws = Cs:

- Allow possible interaction
- Minimize collinearity

Initial E, V, W model

$$\text{logit P}(\mathbf{X}) = \alpha + \beta\text{CAT} + \sum_{i=1}^{5} \gamma_i V_i + \text{CAT} \sum_{j=1}^{5} \delta_j W_j,$$

where V_is = Cs = W_js

However, we have restricted the Vs to the Cs themselves primarily because there are a moderately large number of Cs being considered, and any further addition of Vs is likely to make the model difficult to interpret as well as difficult to fit because of likely collinearity problems.

We next choose the Ws, which are the variables that go into the initial model as product terms with E(CAT). These Ws are the potential effect modifiers to be considered. The Ws that we choose are the Cs themselves, which are also the Vs. That is, W_1 through W_5 equals AGE, CHL, SMK, ECG, and HPT, respectively.

We could have considered other choices for the Ws. For instance, we could have added two-way products of the form W_6 equals AGE × CHL. However, if we added such a term, we would have to add a corresponding two-way product term as a V variable, that is, V_6 equals AGE × CHL, to make our model hierarchically well formulated. This is because AGE × CHL is a lower order component of CAT × AGE × CHL, which is EW_6.

We could also have considered for our set of Ws some subset of the five Cs, rather than all five Cs. For instance, we might have chosen the Ws to be AGE and ECG, so that the corresponding product terms in the model are CAT × AGE and CAT × ECG only.

Nevertheless, we have chosen the Ws to be all five Cs so as to consider the possibility of interaction from any of the five Cs, yet to keep the model relatively small to minimize potential collinearity problems.

Thus, at the end of the variable specification stage, we have chosen as our initial model, the E, V, W model shown here. This model is written in logit form as logit P($\mathbf{X}$) equals a constant term plus terms involving the main effects of the five control variables plus terms involving the interaction of each control variable with the exposure variable CAT.

EXAMPLE (continued)

HWF model?
i.e., given variable, are lower order components in model?

According to our strategy, it is necessary that our initial model, or any subsequently determined reduced model, be hierarchically well formulated. To check this, we assess whether all lower order components of any variable in the model are also in the model.

e.g., CAT × AGE
$\Downarrow$
CAT and AGE both in model as main effects

HWF model? *YES*

For example, the lower order components of a product variable like CAT × AGE are CAT and AGE, and both these terms are in the model as main effects. If we identify the lower order components of any other variable, we can see that the model we are considering is truly hierarchically well formulated.

If CAT × ECG × SMK in model, then *not* HWF model
because
ECG × SMK not in model

Note that if we add to the above model the three-way product term CAT × ECG × SMK, the resulting model is not hierarchically well formulated. This is because the term ECG × SMK has not been specified as one of the *V* variables in the model.

Next
Hierarchical backward elimination procedure

At this point in our model strategy, we are ready to consider simplifying our model by eliminating unnecessary interaction and/or confounding terms. We do this using a hierarchical backward elimination procedure, which considers eliminating the highest-order terms first, then the next highest-order terms, and so on.

First, eliminate *EW* terms
Then, eliminate *V* terms

Because the highest-order terms in our initial model are two-way products of the form *EW*, we first consider eliminating some of these interaction terms. We then consider eliminating the *V* terms, which are the potential confounders.

Interaction assessment
and
confounding assessments (details in Chap. 7)

Here, we summarize the results of the interaction assessment and confounding assessment stages and then return to provide more details of this example in Chap. 7.

Results of Interaction Stage:
CAT × CHL and CAT × HPT
are the only two interaction terms to remain in the model

The results of the interaction stage allow us to eliminate three interaction terms, leaving in the model the two product terms CAT × CHL and CAT × HPT.

Model contains

$$\underbrace{\text{CAT, AGE, CHL, SMK, ECG, HPT,}}_{V\text{s}}$$

CAT × CHL and CAT × HPT

Thus, at the end of interaction assessment, our remaining model contains our exposure variable CAT, the five *V*s namely, AGE, CHL, SMK, ECG, and HPT plus two product terms CAT × CHL and CAT × HPT.

EXAMPLE (continued)

All five *V*s in model so far

Hierarchy principle
identify *V*s that *cannot* be eliminated

$$EV_i \text{ significant}$$
$$\Downarrow$$
$$E \text{ and } V_i \text{ must remain}$$

CAT × CHL ⇒ CAT and CHL remain
CAT × HPT ⇒ CAT and HPT remain

Thus,
CAT (exposure) remains
plus
CHL and HPT remain

AGE, SMK, ECG
eligible for elimination

Results (details in Chap. 7):

Cannot remove AGE, SMK, ECG
(decisions too subjective)

Final model variables:
CAT, AGE, CHL, SMK, ECG, HPT,
CAT × CHL, and CAT × HPT

The reason why the model contains all five *V*s at this point is because we have not yet done any analysis to evaluate which of the *V*s can be eliminated from the model.

However, because we have found two significant interaction terms, we need to use the hierarchy principle to identify certain *V*s that cannot be eliminated from any further models considered.

The hierarchy principle says that all lower order components of significant product terms must remain in all further models.

In our example, the lower order components of CAT × CHL are CAT and CHL, and the lower order components of CAT × HPT are CAT and HPT. Now the CAT variable is our exposure variable, so we will leave CAT in all further models regardless of the hierarchy principle. In addition, we see that CHL and HPT must remain in all further models considered.

This leaves the *V* variables AGE, SMK, and ECG as still being eligible for elimination at the confounding stage of the strategy.

As we show in Chap. 7, we will not find sufficient reason to remove any of the above three variables as nonconfounders. In particular, we will show that decisions about confounding for this example are too subjective to allow us to drop any of the three *V* terms eligible for elimination.

Thus, as a result of our modeling strategy, the final model obtained contains the variables CAT, AGE, CHL, SMK, ECG, and HPT as main effect variables, and it contains the two product terms CAT × CHL and CAT × HPT.

EXAMPLE (continued)

Printout

```
          Variable  Coefficient   S.E.   Chi sq     P
         Intercept    -4.0497    1.2550  10.41   0.0013
               CAT   -12.6894    3.1047  16.71   0.0000
             ┌ AGE     0.0350    0.0161   4.69   0.0303
             │ CHL    -0.00545   0.0042   1.70   0.1923
         Vs <  ECG     0.3671    0.3278   1.25   0.2627
             │ SMK     0.7732    0.3273   5.58   0.0181
             └ HPT     1.0466    0.3316   9.96   0.0016
               ┌ CH   -2.3318    0.7427   9.86   0.0017
             ↗ └ CC    0.0692    0.3316  23.20   0.0000
```

The computer results for this final model are shown here. This includes the estimated regression coefficients, corresponding standard errors, and Wald test information. The variables CAT × HPT and CAT × CHL are denoted in the printout as CH and CC, respectively.

interaction

CH = CAT × HPT and

CC = CAT × CHL

$$\widehat{ROR} = \exp(-12.6894 + 0.0692CHL - 2.3881HPT)$$

Details in Chap. 7.

Also provided here is the formula for the estimated adjusted odds ratio for the CAT, CHD relationship. Using this formula, one can compute point estimates of the odds ratio for different specifications of the effect modifiers CHL and HPT. Further details of these results, including confidence intervals, will be provided in Chap. 7.

SUMMARY

Three stages:
(1) Variable specification
(2) Interaction
(3) Confounding/precision

As a summary of this presentation, we have recommended a modeling strategy with three stages: (1) *variable specification*, (2) *interaction assessment*, and (3) *confounding assessment* followed by consideration of *precision*.

Initial model: HWF model

The initial model has to be *hierarchically well formulated* (HWF). This means that the model must contain all lower order components of any term in the model.

Hierarchical backward elimination procedure
(test for interaction, but do not test for confounding)

Given an initial model, the recommended strategy involves a *hierarchical backward elimination procedure* for removing variables. In carrying out this strategy, statistical testing is allowed for interaction terms, but not for confounding terms.

Hierarchy principle

significant product term
⇓
retain lower order components

When assessing interaction terms, the *hierarchy principle* needs to be applied for any product term found significant. This principle requires all lower order components of significant product terms to remain in all further models considered.

Chapters up to this point

1. Introduction
2. Special Cases

•
•
•

✓ 6. Modeling Strategy Guidelines

7. Strategy for Assessing Interaction and Confounding

This presentation is now complete. We suggest that the reader review the presentation through the detailed outline on the following pages. Then, work through the practice exercises and then the test.

The next chapter is entitled: "Modeling Strategy for Assessing Interaction and Confounding". This continues the strategy described here by providing a detailed description of the interaction and confounding assessment stages of our strategy.

Detailed Outline

I. **Overview** (page 168)
Focus:
- Guidelines for "best" model
- 3-stage strategy
- Valid estimate of *E–D* relationship

II. **Rationale for a modeling strategy** (pages 168–169)
   A. Insufficient explanation provided about strategy in published research; typically only final results are provided.
   B. Too many ad hoc strategies in practice; need for some guidelines.
   C. Need to consider a general strategy that applies to different kinds of modeling procedures.
   D. Goal of strategy in etiologic research is to get a valid estimate of *E–D* relationship; this contrasts with goal of obtaining good prediction, which is built into computer packages for different kinds of models.

III. **Overview of recommended strategy** (pages 169–173)
   A. Three stages: variable specification, interaction assessment, and confounding assessment followed by considerations of precision.
   B. Reason why interaction stage precedes confounding stage: confounding is irrelevant in the presence of strong interaction.
   C. Reason why confounding stage considers precision after confounding is assessed: validity takes precedence over precision.
   D. Statistical concerns needing attention but beyond scope of this presentation: collinearity, controlling the significance level, and influential observations.
   E. The model must be hierarchically well formulated.
   F. The strategy is a hierarchical backward elimination strategy that considers the roles that different variables play in the model and cannot be directly carried out using standard computer algorithms.
   G. Confounding is not assessed by statistical testing.
   H. If interaction is present, confounding assessment is difficult in practice.

IV. **Variable specification stage** (pages 173–175)
   A. Start with *D, E,* and $C_1, C_2, \ldots, C_p$.
   B. Choose *V*s from *C*s based on prior research or theory and considering potential statistical

problems, e.g., collinearity; simplest choice is to let *V*s be *C*s themselves.

C. Choose *W*s from *C*s to be either *V*s or product of two *V*s; usually recommend *W*s to be *C*s themselves or some subset of *C*s.

V. **Causal diagrams** (pages 175–179)

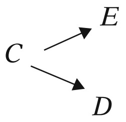
A. The approach for variable selection should consider causal structure.

B. Example of causal diagram: Smoking → Lung Cancer → Abnormal Chest X-ray.

C. Controlling for X-ray status in above diagram leads to bias

D. Depending on the underlying causal structure, adjustment may either remove bias, lead to bias, or be appropriate.

E. Causal diagram for confounding: *C* is a common cause of *E* and *D*



C is a *common cause* of E and D;
The path E–C–D is a (noncausal) *backdoor path* from E to D

F. Other types of causal diagrams:

i. *F* is a common effect of *E* and *D*; conditioning on *F* creates bias; Berkson's bias example.

ii. Example involving unmeasured factors.

G. Conditioning on a common cause can remove bias, whereas conditioning on a common effect can cause bias.

VI. **Other considerations for variable specification** (pages 180–181)

A. Quality of the data: measurement error or misclassification?

B. Qualitative collinearity, e.g., redundant covariates.

C. Sample size

D. Complexity vs. simplicity

E. Know your data! Perform descriptive analyses.

VII. **Hierarchically well-formulated models** (pages 181–184)

A. Definition: given any variable in the model, all lower order components must also be in the model.

B. Examples of models that are and are not hierarchically well-formulated.

C. Rationale: If the model is not hierarchically well-formulated, then tests for significance of the highest-order variables in the model may change with the coding of the variables tested; such tests should be independent of coding.

VIII. The hierarchical backward elimination approach (pages 184–185)

A. Flow diagram representation.

B. Flow description: evaluate EV_iV_j terms first, then EV_i terms, then V_i terms last.

C. Use statistical testing for interaction terms, but decisions about V_i terms should not involve testing.

IX. The hierarchy principle for retaining variables (pages 185–187)

A. Definition: If a variable is to be retained in the model, then all lower order components of that variable are to be retained in the model forever; examples.

C. Rationale: Tests about lower order components can give different conclusions depending on the coding of variables tested; such tests should be independent of coding to be valid; therefore, no such tests are appropriate; example.

X. An example (pages 188–192)

A. Evans County CHD data description.

B. Variable specification stage.

C. Final results.

XI. Summary (pages 192–193)

Practice Exercises

A prevalence study of predictors of surgical wound infection in 265 hospitals throughout Australia collected data on 12,742 surgical patients (McLaws et al., 1988). For each patient, the following independent variables were determined: type of hospital (public or private), size of hospital (large or small), degree of contamination of surgical site (clean or contaminated), and age and sex of the patient. A logistic model was fitted to these data to predict whether or not the patient developed a surgical wound infection during hospitalization. The abbreviated variable names and the manner in which the variables were coded in the model are described as follows:

Variable	Abbreviation	Coding
Type of hospital	HT	1 = public, 0 = private
Size of hospital	HS	1 = large, 0 = small
Degree of contamination	CT	1 = contaminated, 0 = clean
Age	AGE	Continuous
Sex	SEX	1 = female, 0 = male

In the questions that follow, we assume that type of hospital (HT) is considered the exposure variable, and the other four variables are risk factors for surgical wound infection to be considered for control.

1. In defining an *E, V, W* model to describe the effect of HT on the development of surgical wound infection, describe how you would determine the *V* variables to go into the model. (In answering this question, you need to specify the criteria for choosing the *V* variables, rather than the specific variables themselves.)
2. In defining an *E, V, W* model to describe the effect of HT on the development of surgical wound infection, describe how you would determine the *W* variables to go into the model. (In answering this question, you need to specify the criteria for choosing the *W* variables, rather than specifying the actual variables.)
3. State the logit form of a hierarchically well-formulated *E, V, W* model for the above situation in which the *V*s and the *W*s are the *C*s themselves. Why is this model hierarchically well formulated?
4. Suppose the product term HT × AGE × SEX is added to the model described in Exercise 3. Is this new model still hierarchically well formulated? If so, state why; if not, state why not.
5. Suppose for the model described in Exercise 4 that a Wald test is carried out for the significance of the three-factor product term HT × AGE × SEX. Explain

what is meant by the statement that the test result depends on the coding of the variable HT. Should such a test be carried out? Explain briefly.

6. Suppose for the model described in Exercise 3 that a Wald test is carried out for the significance of the two-factor product term HT × AGE. Is this test dependent on coding? Explain briefly.
7. Suppose for the model described in Exercise 3 that a Wald test is carried out for the significance of the main effect term AGE. Why is this test inappropriate here?
8. Using the model of Exercise 3, describe briefly the hierarchical backward elimination procedure for determining the best model.
9. Suppose the interaction assessment stage for the model of Example 3 finds the following two-factor product terms to be significant: HT × CT and HT × SEX; the other two-factor product terms are not significant and are removed from the model. Using the hierarchy principle, what variables must be retained in all further models considered. Can these (latter) variables be tested for significance? Explain briefly.
10. Based on the results in Exercise 9, state the (reduced) model that is left at the end of the interaction assessment stage.

Test

True or False? (Circle T or F)

T F 1. The three stages of the modeling strategy described in this chapter are interaction assessment, confounding assessment, and precision assessment.

T F 2. The assessment of interaction should precede the assessment of confounding.

T F 3. The assessment of interaction may involve statistical testing.

T F 4. The assessment of confounding may involve statistical testing.

T F 5. Getting a precise estimate takes precedence over getting an unbiased answer.

T F 6. During variable specification, the potential confounders should be chosen based on analysis of the data under study.

T F 7. During variable specification, the potential effect modifiers should be chosen by considering prior research or theory about the risk factors measured in the study.

T F 8. During variable specification, the potential effect modifiers should be chosen by considering possible statistical problems that may result from the analysis.

T F 9. A model containing the variables E, A, B, C, A^2, $A \times B$, $E \times A$, $E \times A^2$, $E \times A \times B$, and $E \times C$ is hierarchically well formulated.

T F 10. If the variables $E \times A^2$ and $E \times A \times B$ are found to be significant during interaction assessment, then a *complete* list of all components of these variables that must remain in any further models considered consists of E, A, B, $E \times A$, $E \times B$, and A^2.

The following questions consider the use of logistic regression on data obtained from a matched case-control study of cervical cancer in 313 women from Sydney, Australia (Brock et al., 1988). The outcome variable is cervical cancer status ($1 = \text{present}$, $0 = \text{absent}$). The matching variables are age and socioeconomic status. Additional independent variables not matched on are smoking status, number of lifetime sexual partners, and age at first sexual intercourse. The independent variables are listed below together with their computer abbreviation and coding scheme.

Variable	Abbreviation	Coding
Smoking status	SMK	1 = ever, 0 = never
Number of sexual partners	NS	1 = 4+, 0 = 0–3
Age at first intercourse	AS	1 = 20+, 0 = <19
Age of subject	AGE	Category matched
Socioeconomic status	SES	Category matched

11. Consider the following E, V, W model that considers the effect of smoking, as the exposure variable, on cervical cancer status, controlling for the effects of the other four independent variables listed:

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha &+ \beta\text{SMK} + \textstyle\sum \gamma_i^* V_i^* + \gamma_1\text{NS} + \gamma_2\text{AS}\\ &+ \gamma_3\text{NS} \times \text{AS} + \delta_1\text{SMK} \times \text{NS}\\ &+ \delta_2\text{SMK} \times \text{AS} + \delta_3\text{SMK} \times \text{NS} \times \text{AS},\end{aligned}$$

where the V_i^* are dummy variables indicating matching strata and the γ_i^* are the coefficients of the V_i^* variables. Is this model hierarchically well formulated? If so, explain why; if not, explain why not.

12. For the model in Question 11, is a test for the significance of the three-factor product term SMK × NS × AS dependent on the coding of SMK? If so, explain why; if not explain, why not.

13. For the model in Question 11, is a test for the significance of the two-factor product term SMK × NS dependent on the coding of SMK? If so, explain why; if not, explain why not.
14. For the model in Question 11, briefly describe a hierarchical backward elimination procedure for obtaining a best model.
15. Suppose that the three-factor product term SMK × NS × AS is found significant during the interaction assessment stage of the analysis. Then, using the hierarchy principle, what other *interaction* terms must remain in any further model considered? Also, using the hierarchy principle, what *potential confounders* must remain in any further models considered?
16. Assuming the scenario described in Question 15 (i.e., SMK × NS × AS is significant), what (reduced) model remains after the interaction assessment stage of the model? Are there any potential confounders that are still eligible to be dropped from the model? If so, which ones? If not, why not?

Answers to Practice Exercises

1. The *V* variables should include the *C* variables HS, CT, AGE, and SEX and any functions of these variables that have some justification based on previous research or theory about risk factors for surgical wound infection. The simplest choice is to choose the *V*s to be the *C*s themselves, so that at least every variable already identified as a risk factor is controlled in the simplest way possible.
2. The *W* variables should include some subset of the *V*s, or possibly all the *V*s, plus those functions of the *V*s that have some support from prior research or theory about effect modifiers in studies of surgical wound infection. Also, consideration should be given, when choosing the *W*s, of possible statistical problems, e.g., collinearity, that may arise if the size of the model becomes quite large and the variables chosen are higher order product terms. Such statistical problems may be avoided if the *W*s chosen do not involve very high-order product terms and if the number of *W*s chosen is small. A safe choice is to choose the *W*s to be the *V*s themselves or a subset of the *V*s.
3. $\text{logit P}(\mathbf{X}) = \alpha + \beta\text{HT} + \gamma_1\text{HS} + \gamma_2\text{CT} + \gamma_3\text{AGE}$ $+ \gamma_4 \text{ SEX} + \delta_1\text{HT} \times \text{HS} + \delta_2\text{HT} \times \text{CT}$ $+ \delta_3\text{HT} \times \text{AGE} + \delta_4\text{HT} \times \text{SEX}.$
   This model is HWF because given any interaction term in the model, both of its components are also in the model (as main effects).
4. If HT × AGE × SEX is added to the model, the new model will *not* be hierarchically well formulated because the lower order component AGE × SEX is not contained in the original nor new model.
5. A test for HT × AGE × SEX in the above model is dependent on coding in the sense that different test results (e.g., rejection vs. nonrejection of the null hypothesis) may be obtained depending on whether HT is coded as (0, 1) or (−1, 1) or some other coding. Such a test should not be carried out because any test of interest should be independent of coding, reflecting whatever the real effect of the variable is.
6. A test for HT × AGE in the model of Exercise 3 is independent of coding because the model is hierarchically well formulated and the HT × AGE term is a variable of highest order in the model. (Tests for lower order terms like HT or HS are dependent on the coding even though the model in Exercise 3 is hierarchically well formulated.)
7. A test for the variable AGE is inappropriate because there is a higher order term, HT × AGE, in the model, so that a test for AGE is dependent on the coding of the

HT variable. Such a test is also inappropriate because AGE is a potential confounder, and confounding should not be assessed by statistical testing.

8. A hierarchical backward elimination procedure for the model in Exercise 3 would involve first assessing interaction involving the four interaction terms and then considering confounding involving the four potential confounders. The interaction assessment could be done using statistical testing, whereas the confounding assessment should not use statistical testing. When considering confounding, any *V* variable that is a lower order component of a significant interaction term must remain in all further models and is not eligible for deletion as a nonconfounder. A test for any of these latter *V*s is inappropriate because such a test would be dependent on the coding of any variable in the model.
9. If HT × CT and HT × SEX are found significant, then the *V* variables CT and SEX cannot be removed from the model and must, therefore, be retained in all further models considered. The HT variable remains in all further models considered because it is the exposure variable of interest. CT and SEX are lower order components of higher order interaction terms. Therefore, it is not apropriate to test for their inclusion in the model.
10. At the end of the interaction assessment stage, the remaining model is given by

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha &+ \beta\text{HT} + \gamma_1\text{HS} + \gamma_2\text{CT} + \gamma_3\text{AGE} + \gamma_4\text{SEX} \\ &+ \delta_2\text{HT} \times \text{CT} + \delta_4\text{HT} \times \text{SEX}.\end{aligned}$$

7 Modeling Strategy for Assessing Interaction and Confounding

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_7, © Springer Science+Business Media, LLC 2010

Introduction

This chapter continues the previous chapter (Chap. 6) that gives general guidelines for a strategy for determining a best model using a logistic regression procedure. The focus of this chapter is the interaction and confounding assessment stages of the model building strategy.

We begin by reviewing the previously recommended (Chap. 6) three-stage strategy. The initial model is required to be hierarchically well formulated. In carrying out this strategy, statistical testing is allowed for assessing interaction terms but is not allowed for assessing confounding.

For any interaction term found significant, a hierarchy principle is required to identify lower order variables that must remain in all further models considered. A flow diagram is provided to describe the steps involved in interaction assessment. Methods for significance testing for interaction terms are provided.

Confounding assessment is then described, first when there is no interaction, and then when there is interaction – the latter often being difficult to accomplish in practice.

Finally, an application of the use of the entire recommended strategy is described, and a summary of the strategy is given.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered in this chapter. A detailed outline for review purposes follows the presentation.

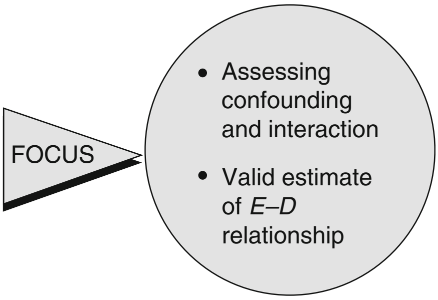
Objectives

Upon completing this chapter, the learner should be able to:

1. Describe and apply the interaction assessment stage in a particular logistic modeling situation.
2. Describe and apply the confounding assessment stage in a particular logistic modeling situation
   a. when there is no interaction and
   b. when there is interaction.

Presentation

I. Overview


FOCUS
• Assessing confounding and interaction
• Valid estimate of *E–D* relationship

This presentation describes a strategy for assessing interaction and confounding when carrying out mathematical modeling using logistic regression. The goal of the strategy is to obtain a valid estimate of an exposure–disease relationship that accounts for confounding and effect modification.

Three stages:

(1) Variable specification
(2) Interaction
(3) Confounding/precision

In the previous presentation on modeling strategy guidelines, we recommended a modeling strategy with three stages: (1) *variable specification*, (2) *interaction assessment*, and (3) *confounding assessment* followed by consideration of *precision*.

Initial model: HWF

The initial model is required to be *hierarchically well formulated*, which we denote as HWF. This means that the initial model must contain all lower order components of any term in the model.

EV_iV_j in initial model $\rightarrow$ $EV_i, EV_j, V_i, V_j, V_iV_j$ also in model

Thus, for example, if the model contains an interaction term of the form EV_iV_j, this will require the lower order terms EV_i, EV_j, V_i, V_j, and V_iV_j also to be in the initial model.

Hierarchical backward elimination:

- Can test for interaction, *but* not confounding
- Can eliminate lower order term if corresponding higher order term is not significant

Given an initial model that is HWF, the recommended strategy then involves a *hierarchical backward elimination procedure* for removing variables. In carrying out this strategy, statistical testing is allowed for interaction terms but not for confounding terms. Note that although any lower order component of a higher order term must belong to the initial HWF model, such a component might be dropped from the model eventually if its corresponding higher order term is found to be nonsignificant during the backward elimination process.

Hierarchy Principle:

Significant product term → All lower order components remain

If, however, when assessing interaction, a product term is found significant, the *Hierarchy Principle* must be applied for lower order components. This principle requires all lower order components of significant product terms to remain in *all* further models considered.

II. Interaction Assessment Stage

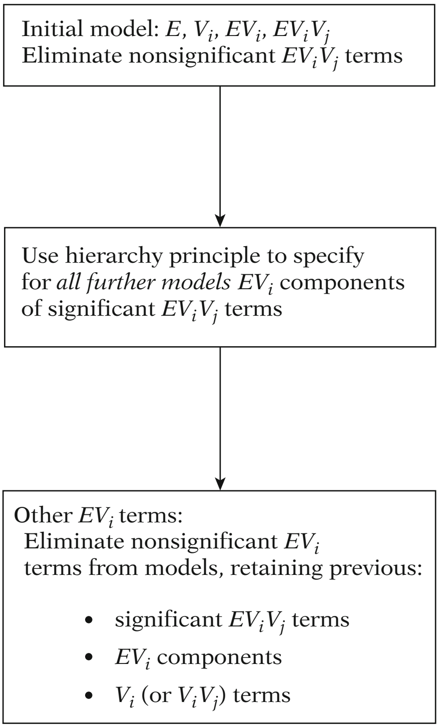
Start with HWF model

Use hierarchical backward elimination:

EV_iV_j before EV_i

Interaction stage flow:

According to our strategy, we consider interaction after we have specified our initial model, which must be hierarchically well formulated (HWF). To address interaction, we use a hierarchical backward elimination procedure, treating higher order terms of the form EV_iV_j prior to considering lower order terms of the form EV_i.



Initial model: E, V_i, EV_i, EV_iV_j
Eliminate nonsignificant EV_iV_j terms

Use hierarchy principle to specify for *all further models* EV_i components of significant EV_iV_j terms

Other EV_i terms:
Eliminate nonsignificant EV_i terms from models, retaining previous:

- significant EV_iV_j terms
- EV_i components
- V_i (or V_iV_j) terms

A flow diagram for the interaction stage is presented here. If our initial model contains terms up to the order EV_iV_j, elimination of these latter terms is considered first. This can be achieved by statistical testing in a number of ways, which we discuss shortly.

When we have completed our assessment of EV_iV_j terms, the next step is to use the hierarchy principle to specify any EV_i terms that are components of significant EV_iV_j terms. Such EV_i terms are to be retained in all further models considered.

The next step is to evaluate the significance of EV_i terms other than those identified by the hierarchy principle. Those EV_i terms that are *nonsignificant* are eliminated from the model. For this assessment, previously significant EV_iV_j terms, their EV_i components, and all V_i terms are retained in any model considered. Note that some of the V_i terms will be of the form V_iV_j if the initial model contains EV_iV_j terms.

Statistical testing
Chunk test for entire collection of interaction terms

In carrying out statistical testing of interaction terms, we recommend that a single "chunk" test for the entire collection (or "chunk") of interaction terms of a given order be considered first.

EXAMPLE

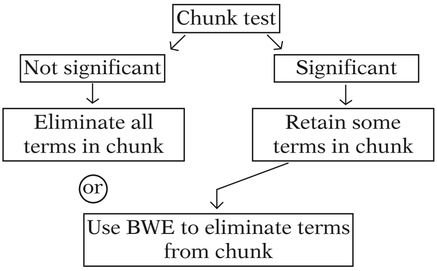
EV_1V_2, EV_1V_3, EV_2V_3 in model

chunk test for $H_0: \delta_1 = \delta_2 = \delta_3 = 0$

use LR statistic $\sim \chi^2_3$ comparing

full model: all V_i, V_iV_j, EV_j, EV_iV_j, E
with reduced model: V_i, V_iV_j, EV_j, E

For example, if there are a total of three EV_iV_j terms in the initial model, namely, EV_1V_2, EV_1V_3, and EV_2V_3, then the null hypothesis for this chunk test is that the coefficients of these variables, say δ_1, δ_2, and δ_3 are all equal to zero. The test procedure is a likelihood ratio (LR) test involving a chi-square statistic with three degrees of freedom, which compares the full model containing all V_i, V_iV_j, EV_i, and EV_iV_j terms with a reduced model containing only V_i, V_iV_j, and EV_i terms, with E in both models.


Chunk test
Not significant
Significant
Eliminate all terms in chunk
Retain some terms in chunk
or
Use BWE to eliminate terms from chunk

If the chunk test *is not significant*, then the investigator may decide to eliminate from the model all terms tested in the chunk, for example, all EV_iV_j terms. If the chunk test *is significant*, then this means that some, but not necessarily all terms in the chunk, are significant and must be retained in the model.

To determine which terms are to be retained, the investigator may carry out a backward elimination (BWE) algorithm to eliminate insignificant variables from the model one at a time. Depending on the preferencen of the investigator, such a BWE procedure may be carried out without even doing the chunk test or regardless of the results of the chunk test.

Even if chunk test n.s.:

- Perform BWE
- May find highly signif. product term(s)
- If so, retain such terms

Alternatively, BWE may still be considered even if the chunk test is nonsignificant. It is possible that one or more product terms are highly significant during BWE, and, if so, should be retained.

EXAMPLE

HWF model:

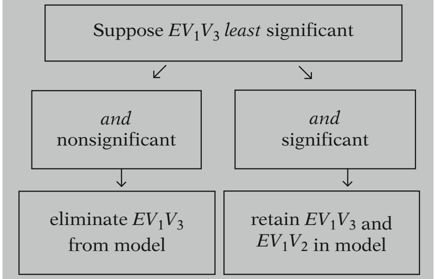
$\boxed{EV_1V_2, EV_1V_3,}$

V_1, V_2, V_3, V_1V_2, V_1V_3,
EV_1, EV_2, EV_3

As an example of such a backward algorithm, suppose we again consider a hierarchically well-formulated model that contains the two EV_iV_j terms EV_1V_2 and EV_1V_3 in addition to the lower order components V_1, V_2, V_3, V_1V_2, V_1V_3, and EV_1, EV_2, EV_3.

EXAMPLE (continued)

BWE approach:


Suppose EV_1V_3 *least* significant
and nonsignificant
and significant
eliminate EV_1V_3 from model
retain EV_1V_3 and EV_1V_2 in model

Using the BWE approach, the least significant EV_iV_j term, say EV_1V_3, is eliminated from the model first, provided it is nonsignificant, as shown on the left-hand side of the flow. If it is significant, as shown on the right-hand side of the flow, then both EV_1V_3 and EV_1V_2 must remain in the model, as do all lower order components, and the modeling process is complete.

Suppose EV_1V_3 not significant:
then drop EV_1V_3 from model.
Reduced model:
EV_1V_2
V_1, V_2, V_3, V_1V_2, V_1V_3
EV_1, EV_2, EV_3

EV_1V_2 dropped if non-signif.

Suppose that the EV_1V_3 term is not significant. Then, this term is dropped from the model. A reduced model containing the remaining EV_1V_2 term and all lower order components from the initial model are then fitted. The EV_1V_2 term is then dropped if nonsignificant but is retained if significant.

Suppose EV_1V_2 significant:
then EV_1V_2 retained and above reduced model is current model

Next: eliminate EV terms

Suppose the EV_1V_2 term is found significant, so that as a result of backward elimination, it is the only three-factor product term retained. Then the above reduced model is our current model, from which we now work to consider eliminating EV terms.

From hierarchy principle:
E, V_1, V_2, EV_1, EV_2, and V_1V_2 retained *in all further models*

Because our reduced model contains the significant term EV_1V_2, we must require (using the hierarchy principle) that the lower order components E, V_1, V_2, EV_1, EV_2, and V_1V_2 are retained in all further models considered.

Assess other EV_i terms:
only EV_3 eligible for removal

The next step is to assess the remaining EV_i terms. In this example, there is only one EV_i term eligible to be removed, namely EV_3, because EV_1 and EV_2 are retained from the hierarchy principle.

EXAMPLE (continued)

LR statistic $\sim \chi^2_1$

Full model: EV_1V_2, EV_1, EV_2, $\boxed{EV_3}$, V_1, V_2, V_3, V_1V_2, V_1V_3

Reduced model: EV_1V_2, EV_1, EV_2, V_1, V_2, V_3, V_1V_2, V_1V_3

$$\text{Wald test}: Z = \frac{\hat{\delta}_{EV_3}}{S_{\hat{\delta}_{EV_3}}}$$

To evaluate whether EV_3 is significant, we can perform a likelihood ratio (LR) chi-square test with one degree of freedom. For this test, the two models being compared are the full model consisting of EV_1V_2, all three EV_i terms and all V_i terms, including those of the form V_iV_j, and the reduced model that omits the EV_3 term being tested. Alternatively, a Wald test can be performed using the Z statistic equal to the coefficient of the EV_3 term divided by its standard error.

Suppose both LR and Wald tests are nonsignificant:
then drop EV_3 from model

Suppose that both the above likelihood ratio and Wald tests are nonsignificant. Then we can drop the variable EV_3 from the model.

Interaction stage results:

EV_1V_2, EV_1, EV_2

$$\left.\begin{matrix} V_1, V_2, V_3 \\ V_1V_2, V_1V_3 \end{matrix}\right\} \text{confounders}$$

Thus, at the end of the interaction assessment stage for this example, the following terms remain in the model: EV_1V_2, EV_1, EV_2, V_1, V_2, V_3, V_1V_2, and V_1V_3.

All V_i (and V_iV_j) remain in model after interaction assessment

All of the V terms, including V_1V_2 and V_1V_3, in the initial model are still in the model at this point. This is because we have been assessing interaction only, whereas the V_1V_2 and V_1V_3 terms concern confounding. Note that although the V_iV_j terms are products, they are potential confounders in this model because they do not involve the exposure variable E.

Most situations
use *only* EV_i
product terms
⇓
interaction assessment
less complicated
⇓
do not need V_iV_j terms
for HWF model.

Before discussing confounding, we point out that for most situations, the highest-order interaction terms to be considered are two-factor product terms of the form EV_i. In this case, interaction assessment begins with such two-factor terms and is often much less complicated to assess than when there are terms of the form EV_iV_j.

In particular, when only two-factor interaction terms are allowed in the model, then it is *not* necessary to have two-factor confounding terms of the form V_iV_j in order for the model to be hierarchically well formulated. This makes the assessment of confounding a less complicated task than when three-factor interactions are allowed.

III. Confounding and Precision Assessment When No Interaction

Confounding:

No statistical testing
(validity issue)

The final stage of our strategy concerns the assessment of confounding followed by consideration of precision. We have previously pointed out that this stage, in contrast to the interaction assessment stage, is carried out without the use of statistical testing. This is because confounding is a validity issue and, consequently, does not concern random error issues that characterize statistical testing.

Confounding	before	precision
↓		↓
Gives correct answer		Gives narrow confidence interval

We have also pointed out that controlling for confounding takes precedence over achieving precision because the primary goal of the analysis is to obtain the correct estimate rather than a narrow confidence interval around the wrong estimate.

No interaction model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \sum \gamma_i V_i$$

(no terms of form EW)

In this section, we focus on the assessment of confounding when the model contains no interaction terms. The model in this case contains only E and V terms but does not contain product terms of the form E times W.

Interaction present?	Confounding assessment?
No	Straightforward
Yes	Difficult

The assessment of confounding is relatively straight-forward when no interaction terms are present in one's model. In contrast, as we shall describe in the next section, it becomes difficult to assess confounding when interaction is present.

EXAMPLE

Initial model

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \cdots + \gamma_5 V_5$$

$$\widehat{\text{OR}} = e^{\hat{\beta}}$$

(a single number)
adjusts for $V_1, \ldots, V_5$

In considering the no interaction situation, we first consider an example involving a logistic model with a dichotomous E variable and five V variables, namely, V_1 through V_5.

For this model, the estimated odds ratio that describes the exposure–disease relationship is given by the expression e to the $\hat{\beta}$, where $\hat{\beta}$ is the estimated coefficient of the E variable. Because the model contains no interaction terms, this odds ratio estimate is a single number that represents an adjusted estimate that controls for all five V variables.

EXAMPLE (continued)

Gold standard estimate:
Controls for all potential confounders (i.e., all five Vs)

Other OR estimates:
Drop some Vs
e.g., drop V_3, V_4, V_5
Reduced model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2$$

$$\widehat{\text{OR}} = e^{\hat{\beta}}$$

controls for V_1 and V_2 only

Reduced model $\neq$ gold standard model

$\boxed{\text{correct answer}}$

$\widehat{\text{OR}}$ (reduced) $\overset{?}{=}$ $\widehat{\text{OR}}$ (gold standard)

If different, then reduced model *does not* control for confounding

Suppose:

Gold standard (all five Vs)
$\widehat{\text{OR}} = 2.5$
reduced model (V_1 and V_2)
↑ $\widehat{\text{OR}} = 5.2$
does not control for confounding
meaningfully different

$$\widehat{\text{OR}}\begin{pmatrix}\text{some other}\\ \text{subset of } V\text{s}\end{pmatrix} \overset{?}{=} \widehat{\text{OR}}\begin{pmatrix}\text{gold}\\ \text{standard}\end{pmatrix}$$

If equal, then subset controls confounding

$\widehat{\text{OR}}$ (V_3 alone) $= 2.7$

$\widehat{\text{OR}}$ (V_4 and V_5) $= 2.3$

$\widehat{\text{OR}}$ (gold standard) $= 2.5$

All three estimates are "essentially" the same as the gold standard

We refer to this estimate as the *gold standard estimate* of effect because we consider it the best estimate we can obtain, which controls for *all* the potential confounders, namely, the five Vs, in our model.

We can nevertheless obtain other estimated odds ratios by dropping some of the Vs from the model. For example, we can drop V_3, V_4, and V_5 from the model and then fit a model containing E, V_1, and V_2. The estimated odds ratio for this "reduced" model is also given by the expression e to the $\hat{\beta}$, where $\hat{\beta}$ is the coefficient of E in the reduced model. This estimate controls for only V_1 and V_2 rather than all five Vs.

Because the reduced model is different from the gold standard model, the estimated odds ratio obtained for the reduced model may be meaningfully different from the gold standard. If so, then we say that the reduced model does not control for confounding because it does not give us the correct answer (i.e., gold standard).

For example, suppose that the gold standard odds ratio controlling for all five Vs is 2.5, whereas the odds ratio obtained when controlling for only V_1 and V_2 is 5.2. Then, because these are meaningfully different odds ratios, we cannot use the reduced model containing V_1 and V_2 because the reduced model does not properly control for confounding.

Now although use of only V_1 and V_2 may not control for confounding, it is possible that some other subset of the Vs may control for confounding by giving essentially the same estimated odds ratio as the gold standard.

For example, perhaps when controlling for V_3 alone, the estimated odds ratio is 2.7 and when controlling for V_4 and V_5, the estimated odds ratio is 2.3. The use of either of these subsets controls for confounding because they give essentially the same answer as the 2.5 obtained for the gold standard.

In general, when no interaction, assess confounding by:

- Monitoring changes in effect measure for subsets of *V*s, i.e., monitor changes in

  $\widehat{\text{OR}} = e^{\hat{\beta}}$

- Identify subsets of *Vs* giving approximately same $\widehat{\text{OR}}$ as gold standard

In general, regardless of the number of *V*s in one's model, the method for assessing confounding when there is no interaction is to monitor changes in the effect measure corresponding to different subsets of potential confounders in the model. That is, we must see to what extent the estimated odds ratio given by e to the $\hat{\beta}$ for a given subset is different from the gold standard odds ratio.

More specifically, to assess confounding, we need to identify subsets of the *V*s that give approximately the same odds ratio as the gold standard. Each of these subsets controls for confounding.

If $\widehat{\text{OR}}$ (subset of *V*s) = $\widehat{\text{OR}}$ (gold standard), then

- which subset to use?
- why not use gold standard?

If we find one or more subsets of the *V*s, which give us the same point estimate as the gold standard, how then do we decide which subset to use? Moreover, why do not we just use the gold standard?

Answer: precision

	less precise	more precise
CIs:	(_____•_____)	(__•__)
	less narrow	more narrow

The answer to both these questions involves consideration of *precision*. By precision, we refer to how narrow a confidence interval around the point estimate is. The narrower the confidence interval, the more precise the point estimate.

EXAMPLE

95% confidence interval (CI)

$\widehat{\text{OR}} = 2.5$ Gold standard all five *V*s	$\widehat{\text{OR}} = 2.7$ Reduced model V_3 only
3.5 – 1.4 = 2.1	4.2 – 1.1 = 3.1
(____•____)	(______•______)
1.4 narrower 3.5	1.1 wider 4.2
more precise	less precise

For example, suppose the 95% confidence interval around the gold standard $\widehat{\text{OR}}$ of 2.5 that controls for all five *V*s has limits of 1.4 and 3.5, whereas the 95% confidence interval around the $\widehat{\text{OR}}$ of 2.7 that controls for V_3 only has limits of 1.1 and 4.2.

Then the gold standard OR estimate is more precise than the OR estimate that controls for V_3 only because the gold standard has the narrower confidence interval. Specifically, the narrower width is 3.5 minus 1.4, or 2.1, whereas the wider width is 4.2 minus 1.1, or 3.1.

CI for GS may be either less precise or more precise than CI for subset

Note that it is possible that the gold standard estimate actually may be less precise than an estimate resulting from control of a subset of *V*s. This will depend on the particular data set being analyzed.

Why do not we use gold standard?

Answer. Might find subset of *V*s that will

- gain precision (narrower CI)
- without sacrificing validity (same point estimate)

The answer to the question why do not we just use the gold standard is that we might gain a meaningful amount of precision controlling for a subset of *V*s without sacrificing validity. That is, we might find a subset of *V*s to give essentially the same estimate as the gold standard but which also has a much narrower confidence interval.

EXAMPLE

Model	$\widehat{OR}$	CI
✓ V_4 and V_5	same (2.3)	narrower (1.9, 3.1)
Gold standard	same (2.5)	wider (1.4, 3.5)

For instance, controlling for V_4 and V_5 may obtain the same point estimate as the gold standard but a narrower confidence interval, as illustrated here. If so, we would prefer the estimate that uses V_4 and V_5 in our model to the gold standard estimate.

Which subset to control?

Answer. subset with most meaningful gain in precision

We also asked the question, "How do we decide which subset to use for control?" The answer to this is to choose that subset which gives the most meaningful gain in precision among all eligible subsets, including the gold standard.

Eligible subset. same point estimate as gold standard

By *eligible subset*, we mean any collection of *V*s that gives essentially the same point estimate as the gold standard.

Recommended procedure:

(1) Identify eligible subsets of *V*s
(2) Control for that subset with largest gain in precision

However, if no subset gives *better* precision, use gold standard

Thus, we recommend the following general procedure for the confounding and precision assessment stage of our strategy:

(1) *Identify eligible subsets* of *V*s giving approximately the same odds ratio as the gold standard.
(2) Control for that subset which gives the largest gain in precision. However, if no subset gives meaningfully better precision than the gold standard, it is *scientifically* better to control for all *V*s using the gold standard.

Scientific: Gold standard uses *all* relevant variables for control

The gold standard is *scientifically* better because persons who critically appraise the results of the study can see that when using the gold standard, all the relevant variables have been controlled for in the analysis.

EXAMPLE

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \cdots + \gamma_5 V_5$$

Vs in model	$e^{\hat{\beta}}$	95% CI
V_1, V_2, V_3, V_4, V_5	2.5	(1.4, 3.5)
V_3 only	2.7	(1.1, 4.2)
V_4, V_5 only	2.3	(1.3, 3.4)
other subsets	*	—

same width

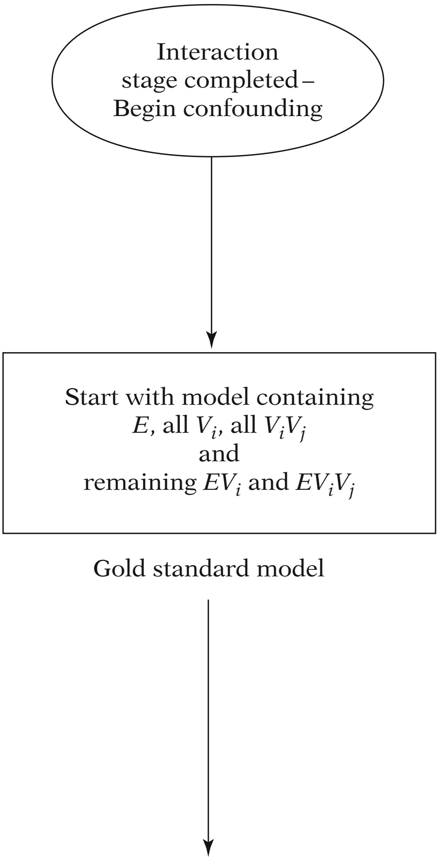
wider

*$e^{\hat{\beta}}$ meaningfully different from 2.5

Returning to our example involving five V variables, suppose that the point estimates and confidence intervals for various subsets of Vs are given as shown here. Then there are only two eligible subsets other than the gold standard – namely V_3 alone, and V_4 and V_5 together because these two subsets give the same odds ratio as the gold standard.

Considering precision, we then conclude that we should control for all five Vs, that is, the gold standard, because no meaningful gain in precision is obtained from controlling for either of the two eligible subsets of Vs. Note that when V_3 alone is controlled, the CI is wider than that for the gold standard. When V_4 and V_5 are controlled together, the CI is the same as the gold standard.

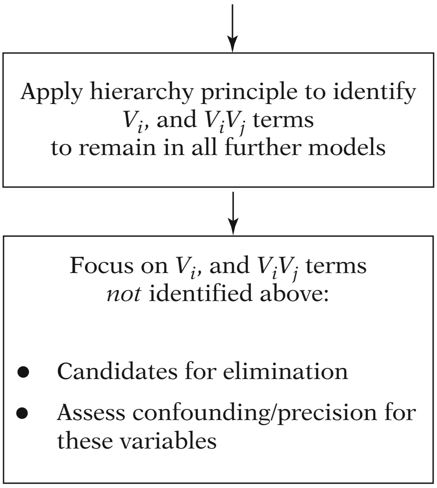
IV. Confounding Assessment with Interaction


Interaction stage completed – Begin confounding
Start with model containing E, all V_i, all V_iV_j and remaining EV_i and EV_iV_j
Gold standard model

We now consider how to assess confounding when the model contains interaction terms. A flow diagram that describes our recommended strategy for this situation is shown here. This diagram starts from the point in the strategy where interaction has already been assessed. Thus, we assume that decisions have been made about which interaction terms are significant and are to be retained in all further models considered.

In the first step of the flow diagram, we start with a model containing E and all potential confounders initially specified as V_i and V_iV_j terms plus remaining interaction terms determined from interaction assessment. This includes those EV_i and EV_iV_j terms found to be significant plus those EV_i terms that are components of significant EV_iV_j terms. Such EV_i terms must remain in all further models considered because of the hierarchy principle.

This model is the *gold standard model* to which all further models considered must be compared. By gold standard, we mean that the odds ratio for this model controls for all potential confounders in our initial model, that is, all the V_is and V_iV_js.



Apply hierarchy principle to identify V_i, and V_iV_j terms to remain in all further models

Focus on V_i, and V_iV_j terms *not* identified above:

- Candidates for elimination
- Assess confounding/precision for these variables

In the second step of the flow diagram, we apply the hierarchy principle to identify those V_i and V_iV_j terms that are lower order components of those interaction terms found significant. Such lower order components must remain in all further models considered.

In the final step of the flow diagram, we focus on only those V_i and V_iV_j terms not identified by the hierarchy principle. These terms are *candidates* to be dropped from the model as nonconfounders. For those variables identified as candidates for elimination, we then *assess confounding followed by* consideration of *precision*.

Interaction terms in model

$\Downarrow$

Final (confounding) step difficult – subjective

If the model contains interaction terms, the final (confounding) step is difficult to carry out and requires subjectivity in deciding which variables can be eliminated as nonconfounders. We will illustrate such difficulties by the example below.

Safest approach:

Keep all potential confounders in model: controls confounding but may lose precision

To avoid making subjective decisions, the safest approach is to keep all potential confounders in the model, whether or not they are eligible to be dropped. This will ensure the proper control of confounding but has the potential drawback of not giving as precise an odds ratio estimate as possible from some smaller subset of confounders.

Confounding – general procedure:

$\widehat{\text{OR}}$ change?

Gold standard model vs. Model without one or more V_i and V_iV_j

(1) Identify subsets so that $\widehat{\text{OR}}_{\text{GS}} \approx \widehat{\text{OR}}_{\text{subset}}$.

(2) Control for largest gain in precision.

Difficult when there is interaction

In assessing confounding when there are interaction terms, the general procedure is analogous to when there is no interaction. We assess whether the estimated odds ratio changes from the gold standard model when compared to a model without one or more of the eligible V_is and V_iV_js.

More specifically, we carry out the following two steps:

(1) Identify those subsets of V_is and V_iV_js giving approximately the same odds ratio estimate as the gold standard (GS).

(2) Control for that subset which gives the largest gain in precision.

Interaction: $\widehat{\text{OR}} = \exp\left(\hat{\beta} + \sum \hat{\delta}_j W_j\right)$

$\hat{\beta}$ and $\hat{\delta}_j$ nonzero

no interaction: $\widehat{\text{OR}} = \exp(\hat{\beta})$

If the model contains interaction terms, the first step is difficult in practice. The odds ratio expression, as shown here, involves two or more coefficients, including one or more nonzero $\hat{\delta}$. In contrast, when there is no interaction, the odds ratio involves the single coefficient $\hat{\beta}$.

Coefficients change when potential confounders dropped:

- Meaningful change?
- Subjective?

It is likely that at least one or more of the $\hat{\beta}$ and $\hat{\delta}$ coefficients will change somewhat when potential confounders are dropped from the model. To evaluate how much of a change is a *meaningful* change when considering the collection of coefficients in the odds ratio formula is quite *subjective*. This will be illustrated by the example.

EXAMPLE

Variables in initial model:

$E, V_1, V_2, V_3, V_4 = V_1V_2$

$EV_1, EV_2, EV_3, EV_4 = EV_1V_2$

As an example, suppose our initial model contains E, four Vs, namely, V_1, V_2, V_3, and $V_4 = V_1V_2$, and four EVs, namely, EV_1, EV_2, EV_3, and EV_4. Note that EV_4 alternatively can be considered as a three-factor product term as it is of the form EV_1V_2.

Suppose EV_4 ($= EV_1\ V_2$) significant

Suppose also that because EV_4 is a three-factor product term, it is tested first, after all the other variables are forced into the model. Further, suppose that this test is significant, so that the term EV_4 is to be retained in all further models considered.

Hierarchy principle:

EV_1 and EV_2 retained in all further models

EV_3 candidate to be dropped

Because of the hierarchy principle, then, we must retain EV_1 and EV_2 in all further models as these two terms are components of EV_1V_2. This leaves EV_3 as the only remaining two-factor interaction candidate to be dropped if not significant.

Test for EV_3 (LR or Wald test)

To test for EV_3, we can do either a likelihood ratio test or a Wald test for the addition of EV_3 to a model after E, V_1, V_2, V_3, $V_4 = V_1V_2$, EV_1, EV_2, and EV_4 are forced into the model.

V_1, V_2, V_3, V_4 (*all* potential confounders) forced into model during interaction stage

Note that all four potential confounders – V_1 through V_4 – are forced into the model here because we are at the interaction stage so far, and we have not yet addressed confounding in this example.

EXAMPLE (continued)

LR test for EV_3: Compare *full model* containing

$$E, \underbrace{V_1, V_2, V_3, V_4}_{V\text{s}}, \underbrace{EV_1, EV_2, EV_3, EV_4}_{EV\text{s}}$$

with *reduced model* containing

$$E, V_1, V_2, V_3, V_4, \underbrace{EV_1, EV_2, EV_4}_{\text{without } EV_3}$$

$$\text{LR} = (-2 \ln \hat{L}_{\text{reduced}}) - (-2 \ln \hat{L}_{\text{full}})$$

is $\chi^2_{1\text{df}}$ under H_0: $\delta_{EV_3} = 0$ *in full model*

Suppose EV_3 *not* significant

⇓

model after interaction assessment:

$$E, \boxed{V_1, V_2, V_3, V_4}, EV_1, EV_2, EV_4$$

where $V_4 = V_1V_2$ potential confounders

Hierarchy principle:
identify Vs not eligible to be dropped – lower order components

EV_1V_2 significant
⇓ Hierarchy principle
Retain V_1, V_2, and $V_4 = V_1V_2$
Only V_3 eligible to be dropped

$$\widehat{\text{OR}}_{V_1,V_2,V_3,V_4} \overset{?}{\neq} \widehat{\text{OR}}_{V_1,V_2,V_4}$$

↑

excludes V_3

The likelihood ratio test for the significance of EV_3 compares a "full" model containing E, the four Vs, EV_1, EV_2, EV_3, and EV_4 with a reduced model that eliminates EV_3 from the full model.

The LR statistic is given by the difference in the log likelihood statistics for the full and reduced models. This statistic has a chi-square distribution with one degree of freedom under the null hypothesis that the coefficient of the EV_3 term is 0 in our full model at this stage.

Suppose that when we carry out the LR test for this example, we find that the EV_3 term is not significant. Thus, at the end of the interaction assessment stage, we are left with a model that contains E, the four Vs, EV_1, EV_2, and EV_4. We are now ready to assess confounding for this example.

Our initial model contained four potential confounders, namely, V_1 through V_4, where V_4 is the product term V_1 times V_2. Because of the hierarchy principle, some of these terms are not eligible to be dropped from the model, namely, the lower order components of higher order product terms remaining in the model.

In particular, because EV_1V_2 has been found significant, we must retain in all further models the lower order components V_1, V_2, and V_1V_2, which equals V_4. This leaves V_3 as the only remaining potential confounder that is eligible to be dropped from the model as a possible nonconfounder.

To evaluate whether V_3 can be dropped from the model as a nonconfounder, we consider whether the odds ratio for the model that controls for all four potential confounders, including V_3, plus previously retained interaction terms, is meaningfully different from the odds ratio that controls for previously retained variables but excludes V_3.

EXAMPLE (continued)

$\widehat{\text{OR}}_{V_1,V_2,V_3,V_4} = \exp\left(\hat{\beta} + \hat{\delta}_1 V_1 + \hat{\delta}_2 V_2 + \hat{\delta}_4 V_4\right)$,

where $\hat{\delta}_1$, $\hat{\delta}_2$, and $\hat{\delta}_4$ are coefficients of EV_1, EV_2, and $EV_4 = EV_1V_2$

The odds ratio that controls for all four potential confounders plus retained interaction terms is given by the expression shown here. This expression gives a formula for calculating numerical values for the odds ratio. This formula contains the coefficients $\hat{\beta}, \hat{\delta}_1, \hat{\delta}_2$, and $\hat{\delta}_4$, but also requires specification of three effect modifiers – namely, V_1, V_2, and V_4, which are in the model as product terms with E.

$\widehat{\text{OR}}$ differs for different specifications of V_1, V_2, V_4

The numerical value computed for the odds ratio will differ depending on the values specified for the effect modifiers V_1, V_2, and V_4. This should not be surprising because the presence of interaction terms in the model means that the value of the odds ratio differs for different values of the effect modifiers.

Gold standard $\widehat{\text{OR}}$:

- Controls for all potential confounders
- Gives baseline $\widehat{\text{OR}}$

The above odds ratio is the *gold standard* odds ratio expression for our example. This odds ratio controls for all potential confounders being considered, and it provides baseline odds ratio values to which all other odds ratio computations obtained from dropping candidate confounders can be compared.

$\widehat{\text{OR}}^* = \exp(\hat{\beta}^* + \hat{\delta}_1^* V_1 + \hat{\delta}_2^* V_2 + \hat{\delta}_4^* V_4)$,

where $\hat{\beta}^*, \hat{\delta}_1^*, \hat{\delta}_2^*, \hat{\delta}_4^*$ are coefficients in model without V_3

The odds ratio that controls for previously retained variables but excludes the control of V_3 is given by the expression shown here. Note that this expression is essentially of the same form as the gold standard odds ratio. In particular, both expressions involve the coefficient of the exposure variable and the same set of effect modifiers.

Model without V_3:

$$E, V_1, V_2, V_4, EV_1, EV_2, EV_4$$

Model with V_3:

$$E,\ V_1,\ V_2, \enclose{circle}{V_3} V_4,\ EV_1,\ EV_2,\ EV_4$$

However, the estimated coefficients for this odds ratio are denoted with an asterisk (*) to indicate that these estimates may differ from the corresponding estimates for the gold standard. This is because the model that excludes V_3 contains a different set of variables and, consequently, may result in different estimated coefficients for those variables in common to both models.

Possible that

$\hat{\beta} \neq \hat{\beta}^*, \hat{\delta}_1 \neq \hat{\delta}_1^*, \hat{\delta}_2 \neq \hat{\delta}_2^*, \hat{\delta}_4 \neq \hat{\delta}_4^*$

In other words, because the gold standard model contains V_3, whereas the model for the asterisked odds ratio does not contain V_3, it is possible that $\hat{\beta}$ will differ from $\hat{\beta}^*$, and that the $\hat{\delta}$ will differ from the $\hat{\delta}^*$.

EXAMPLE (continued)

Meaningful difference?

Gold standard model:

$$\widehat{\mathrm{OR}} = \exp\left(\hat{\beta} + \hat{\delta}_1 V_1 + \hat{\delta}_2 V_2 + \hat{\delta}_4 V_4\right)$$

Model without V_3:

$$\widehat{\mathrm{OR}}* = \exp\left(\hat{\beta}^* + \hat{\delta}_1^* V_1 + \hat{\delta}_2^* V_2 + \hat{\delta}_4^* V_4\right)$$

$$\boxed{\hat{\beta}, \hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_4} \text{ vs. } \boxed{\hat{\beta}^*, \hat{\delta}_1^*, \hat{\delta}_2^*, \hat{\delta}_4^*}$$

Difference?

Yes $\Rightarrow$ V_3 confounder; cannot eliminate V_3

No $\Rightarrow$ V_3 not confounder; drop V_3 if precision gain

Difficult approach:

- Four coefficients to compare
- Coefficients likely to change

Overall decision required about change in

$\hat{\beta}, \hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_4$

More subjective than when no interaction (only $\hat{\beta}$)

To assess (data-based) confounding here, we must determine whether there is a meaningful difference between the gold standard and asterisked odds ratio expressions. There are two alternative ways to do this. (The assessment of confounding involves criteria beyond what may exist in the data.)

One way is to compare corresponding estimated coefficients in the odds ratio expression, and then to make a decision whether there is a meaningful difference in one or more of these coefficients.

If we decide *yes*, that there is a difference, we then conclude that there is confounding due to V_3, so that we cannot eliminate V_3 from the model. If, on the other hand, we decide *no*, that corresponding coefficients are not different, we then conclude that we do not need to control for the confounding effects of V_3. In this case, we may consider dropping V_3 from the model if we can gain precision by doing so.

Unfortunately, this approach for assessing confounding is difficult in practice. In particular, in this example, the odds ratio expression involves four coefficients, and it is likely that at least one or more of these will change somewhat when one or more potential confounders are dropped from the model.

To evaluate whether there is a meaningful change in the odds ratio therefore requires an overall decision as to whether the collection of four coefficients, $\hat{\beta}$ and three $\hat{\delta}$, in the odds ratio expression meaningfully change. This is a more subjective decision than for the no interaction situation when $\hat{\beta}$ is the only coefficient to be monitored.

EXAMPLE (continued)

$$\widehat{\mathrm{OR}} = \exp \underbrace{(\hat{\beta} + \hat{\delta}_1 V_1 + \hat{\delta}_2 V_2 + \hat{\delta}_4 V_4)}_{\text{linear function}}$$

$\hat{\beta}, \hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_4$ on log odds ratio scale, but odds ratio scale is clinically relevant

Log odds ratio scale:

$\hat{\beta} = -12.69$ vs. $\hat{\beta}^* = -12.72$

$\hat{\delta}_1 = 0.0692$ vs. $\hat{\delta}_1^* = 0.0696$

Odds ratio scale:

Calculate $\widehat{\mathrm{OR}} = \exp\left(\hat{\beta} + \sum \hat{\delta}_j W_j\right)$

for different choices of W_j

Gold standard OR:

$\widehat{\mathrm{OR}} = \exp(\hat{\beta} + \hat{\delta}_1 V_1 + \hat{\delta}_2 V_2 + \hat{\delta}_4 V_4)$, where $V_4 = V_1 V_2$.

Specify V_1 and V_2 to get OR:

	$V_1 = 20$	$V_1 = 30$	$V_1 = 40$
$V_2 = 100$	$\widehat{\mathrm{OR}}$	$\widehat{\mathrm{OR}}$	$\widehat{\mathrm{OR}}$
$V_2 = 200$	$\widehat{\mathrm{OR}}$	$\widehat{\mathrm{OR}}$	$\widehat{\mathrm{OR}}$

Model without V_3:

$\widehat{\mathrm{OR}}^* = \exp(\hat{\beta}^* + \hat{\delta}_1^* V_1 + \hat{\delta}_2^* V_2 + \hat{\delta}_4^* V_4)$

	$V_1 = 20$	$V_1 = 30$	$V_1 = 40$
$V_2 = 100$	$\widehat{\mathrm{OR}}$*	$\widehat{\mathrm{OR}}$*	$\widehat{\mathrm{OR}}$*
$V_2 = 200$	$\widehat{\mathrm{OR}}$*	$\widehat{\mathrm{OR}}$*	$\widehat{\mathrm{OR}}$*

Compare tables of

$\widehat{\mathrm{OR}}$s vs. $\widehat{\mathrm{OR}}$*s

gold standard   model without V_3

Moreover, because the odds ratio expression involves the exponential of a linear function of the four coefficients, these coefficients are on a log odds ratio scale rather than an odds ratio scale. Using a log scale to judge the meaningfulness of a change is not as clinically relevant as using the odds ratio scale.

For example, a change in $\hat{\beta}$ from -12.69 to -12.72 and a change in $\hat{\delta}_1$ from 0.0692 to 0.0696 are not easy to interpret as clinically meaningful because these values are on a log odds ratio scale.

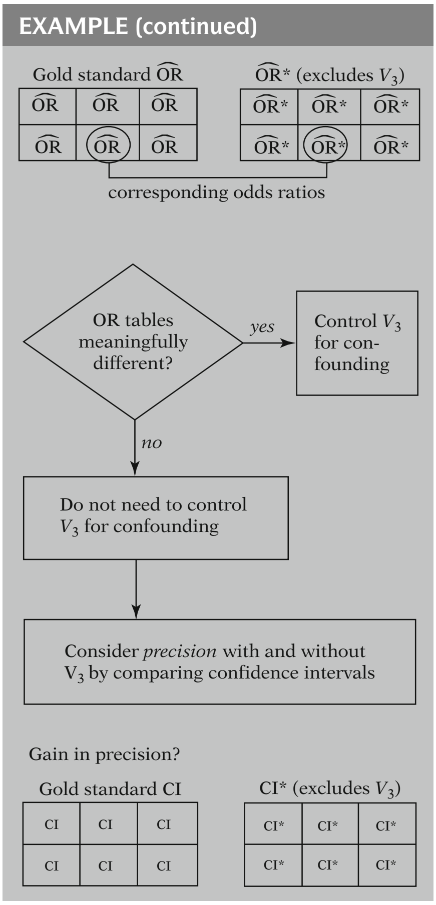
A more interpretable approach, therefore, is to view such changes on the odds ratio scale. This involves calculating numerical values for the odds ratio by substituting into the odds ratio expression different choices of the values for the effect modifiers W_j.

Thus, to calculate an odds ratio value from the gold standard formula shown here, which controls for all four potential confounders, we would need to specify values for the effect modifiers V_1, V_2, and V_4, where V_4 equals $V_1 V_2$. For different choices of V_1 and V_2, we would then obtain different odds ratio values. This information can be summarized in a table or graph of odds ratios, which consider the different specifications of the effect modifiers. A sample table is shown here.

To assess confounding on an odds ratio scale, we would then compute a similar table or graph, which would consider odds ratio values for a model that drops one or more eligible V variables. In our example, because the only eligible variable is V_3, we, therefore, need to obtain an odds ratio table or graph for the model that does not contain V_3. A sample table of OR* values is shown here.

Thus, to assess whether we need to control for confounding from V_3, we need to compare two tables of odds ratios, one for the gold standard and the other for the model that does not contain V_3.

EXAMPLE (continued)

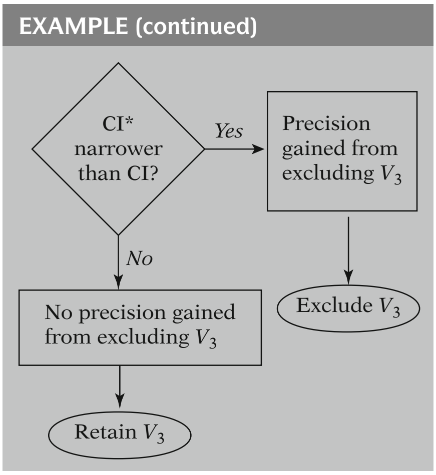

Gold standard $\widehat{OR}$
$\widehat{OR}$* (excludes V_3)
$\widehat{OR}$
$\widehat{OR}$*
corresponding odds ratios
OR tables meaningfully different?
yes
Control V_3 for con-founding
no
Do not need to control V_3 for confounding
Consider *precision* with and without V_3 by comparing confidence intervals
Gain in precision?
Gold standard CI
CI* (excludes V_3)
CI
CI*

If, looking at these two tables collectively, we find that *yes*, there is one or more meaningful difference in corresponding odds ratios, we would conclude that the variable V_3 needs to be controlled for confounding. In contrast, if we decide that *no*, the two tables are not meaningfully different, we can conclude that variable V_3 does not need to be controlled for confounding.

If the decision is made that V_3 does not need to be controlled for confounding reasons, we still may wish to control for V_3 because of precision reasons. That is, we can compare confidence intervals for corresponding odds ratios from each table to determine whether we gain or lose precision depending on whether or not V_3 is in the model.

In other words, to assess whether there is a gain in precision from dropping V_3 from the model, we need to make an overall comparison of two tables of confidence intervals for odds ratio estimates obtained when V_3 is in and out of the model.

EXAMPLE (continued)


CI* narrower than CI?
Yes
Precision gained from excluding V_3
Exclude V_3
No
No precision gained from excluding V_3
Retain V_3

If, overall, we decide that *yes*, the asterisked confidence intervals, which exclude V_3, are narrower than those for the gold standard table, we would conclude that precision is gained from excluding V_3 from the model. Otherwise, if we decide *no*, then we conclude that no meaningful precision is gained from dropping V_3, and so we retain this variable in our final model.

Confounding assessment when interaction present (summary):

- Compare tables of ORs and CIs
- Subjective – debatable
- Safest decision – control for all potential counfounders

Thus, we see that when there is interaction and we want to assess both confounding and precision, we must compare tables of odds ratio point estimates followed by tables of odds ratio confidence intervals. Such comparisons are quite subjective and, therefore, debatable in practice. That is why the safest decision is to control for all potential confounders even if some *V*s are candidates to be dropped.

V. The Evans County Example Continued

EXAMPLE

Evans County Heart Disease Study

$n = 609$ white males
9-year follow-up

$D = \text{CHD}_{(0,\,1)}$
$E = \text{CAT}_{(0,\,1)}$

$C\text{s} : \underbrace{\text{AGE, CHL}}_{\text{continuous}}\ \underbrace{\text{SMK, ECG, HPT}}_{(0,\,1)}$

We now review the interaction and confounding assessment recommendations by returning to the Evans County Heart Disease Study data that we have considered in the previous chapters.

Recall that the study data involves 609 white males followed for 9 years to determine CHD status. The exposure variable is catecholamine level (CAT), and the *C* variables considered for control are AGE, cholesterol (CHL), smoking status (SMK), electrocardiogram abnormality status (ECG), and hypertension status (HPT). The variables AGE and CHL are treated continuously, whereas SMK, ECG, and HPT are (0, 1) variables.

EXAMPLE (continued)

Initial *E, V, W* model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta\text{CAT} + \sum_{i=1}^{5} \gamma_i V_i + E\sum_{j=1}^{5} \delta_j W_j,$$

where *V*s = *C*s = *W*s
HWF model because

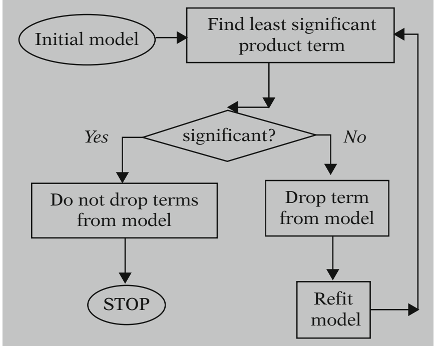
EV_i in model
⇓
E and V_i in model

Highest order in model: EV_i
no EV_iV_j or V_iV_j terms

Next step:
Interaction assessment using backward elimination (BWE)

(Note: Chunk test for
H_0: $\delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = 0$
is highly significant)

Backward elimination (BWE):



Interaction results:

Eliminated	Remaining
CAT × AGE	CAT × CHL
CAT × SMK	CAT × HPT
CAT × ECG	

In the variable specification stage of our strategy, we choose an initial *E, V, W* model, shown here, containing the exposure variable CAT, five *V*s which are the *C*s themselves, and five *W*s which are also the *C*s themselves and which go into the model as product terms with the exposure CAT.

This initial model is HWF because the lower order components of any EV_i term, namely, E and V_i, are contained in the model.

Note also that the highest-order terms in this model are two-factor product terms of the form EV_i. Thus, we are not considering more complicated three-factor product terms of the form EV_iV_j nor V_i terms that are of the form V_iV_j.

The next step in our modeling strategy is to consider eliminating unnecessary interaction terms. To do this, we use a backward elimination (BWE) procedure to remove variables. For interaction terms, we proceed by eliminating (BWE) product terms one at a time.

The flow for our backward procedure begins with the initial model and then identifies the least significant product term. We then ask, "Is this term significant?" If our answer is *no*, we eliminate this term from the model. The model is then refitted using the remaining terms. The least significant of these remaining terms is then considered for elimination.

This process continues until our answer to the significance question in the flow diagram is *yes*. If so, the least significant term is significant in some refitted model. Then, no further terms can be eliminated, and our process must stop.

For our initial Evans County model, the BWE allows us to eliminate the product terms of CAT × AGE, CAT × SMK, and CAT × ECG. The remaining interaction terms are CAT × CHL and CAT × HPT.

EXAMPLE (continued)

Printout:

	Variable	Coefficient	S.E.	Chi sq	*P*
	Intercept	−4.0497	1.2550	10.41	0.0013
	CAT	−12.6894	3.1047	16.71	0.0000
*V*s	AGE	0.0350	0.0161	4.69	0.0303
	CHL	−0.00545	0.0042	1.70	0.1923
	ECG	0.3671	0.3278	1.25	0.2627
	SMK	0.7732	0.3273	5.58	0.0181
	HPT	1.0466	0.3316	9.96	0.0016
	CH	−2.3318	0.7427	9.86	0.0017
	CC	0.0692	0.3316	23.20	0.0000

*W*s

CH = CAT × HPT and CC = CAT × CHL remain in all further models

A summary of the printout for the model remaining after interaction assessment is shown here. In this model, the two interaction terms are CH equals CAT × HPT and CC equals CAT × CHL. The least significant of these two terms is CH because the Wald statistic for this term is given by the chi-square value of 9.86, which is less significant than the chi-square value of 23.20 for the CC term.

The *P*-value for the CH term is 0.0017, so that this term is significant at well below the 1% level. Consequently, we cannot drop CH from the model, so that all further models must contain the two product terms CH and CC.

Confounding assessment:
Step 1. Variables in model:

$$\text{CAT}, \underbrace{\text{AGE, CHL, SMK, ECG, HPT}}_{V\text{s}}$$

$$\underbrace{\text{CAT} \times \text{CHL}, \text{CAT} \times \text{HPT},}_{EV\text{s}}$$

We are now ready to consider the confounding assessment stage of the modeling strategy. The first step in this stage is to identify all variables remaining in the model after the interaction stage. These are CAT, all five *V* variables, and the two product terms CAT × CHL and CAT × HPT.

All five *V*s still in model after interaction

The reason why the model contains all five *V*s at this point is that we have only completed interaction assessment and have not yet begun to address confounding to evaluate which of the *V*s can be eliminated from the model.

Hierarchy principle:

- Determine *V*s that cannot be eliminated
- All lower order components of significant product terms remain

The next step is to apply the hierarchy principle to determine which *V* variables cannot be eliminated from further models considered.

The hierarchy principle requires all lower order components of significant product terms to remain in all further models.

CAT × CHL significant ⇒ CAT and CHL components

CAT × HPT significant ⇒ CAT and HPT components

The two significant product terms in our model are CAT × CHL and CAT × HPT. The lower order components of CAT × CHL are CAT and CHL. The lower order components of CAT × HPT are CAT and HPT.

EXAMPLE (continued)

Thus, retain CAT, CHL, and HPT in all further models

Because CAT is the exposure variable, we must leave CAT in all further models regardless of the hierarchy principle. In addition, CHL and HPT are the two *V*s that must remain in all further models.

Candidates for elimination:
AGE, SMK, ECG

This leaves the *V* variables AGE, SMK, and ECG as still being candidates for elimination as possible nonconfounders.

Assessing confounding:
Do coefficients in $\widehat{\text{OR}}$ expression change?

As described earlier, one approach to assessing whether AGE, SMK, and ECG are nonconfounders is to determine whether the coefficients in the odds ratio expression for the CAT, CHD relationship change meaningfully as we drop one or more of the candidate terms AGE, SMK, and ECG.

$$\widehat{\text{OR}} = \exp(\hat{\beta} + \hat{\delta}_1\text{CHL} + \hat{\delta}_2\text{HPT}),$$

where

$\hat{\beta}$ = coefficient of CAT
$\hat{\delta}_1$ = coefficient of CC = CAT × CHL
$\hat{\delta}_2$ = coefficient of CH = CAT × HPT

The odds ratio expression for the CAT, CHD relationship is shown here. This expression contains $\hat{\beta}$, the coefficient of the CAT variable, plus two terms of the form $\hat{\delta}$ times *W*, where the *W*s are the effect modifiers CHL and HPT that remain as a result of interaction assessment.

Gold standard $\widehat{\text{OR}}$ (all *V*s):

$$\widehat{\text{OR}} = \exp(\hat{\beta} + \hat{\delta}_1\text{CHL} + \hat{\delta}_2\text{HPT}),$$

where

$\hat{\beta} = -12.6894, \hat{\delta}_1 = 0.0692, \hat{\delta}_2 = -2.3318$

The gold standard odds ratio expression is derived from the model remaining after interaction assessment. This model controls for all potential confounders, that is, the *V*s, in the initial model. For the Evans County data, the coefficients in this odds ratio, which are obtained from the printout above, are $\hat{\beta}$ equals −12.6894, $\hat{\delta}_1$ equals 0.0692, and $\hat{\delta}_2$ equals −2.3318.

V_i in model	$\hat{\beta}$	$\hat{\delta}_1$	$\hat{\delta}_2$
All five *V* variables	−12.6894	0.0692	−2.3318
CHL, HPT, AGE, ECG	−12.7285	0.0697	−2.3836
CHL, HPT, AGE, SMK	−12.8447	0.0707	−2.3334
CHL, HPT, ECG, SMK	−12.5684	0.0697	−2.2081
CHL, HPT, AGE	−12.7879	0.0707	−2.3796
CHL, HPT, ECG	−12.6850	0.0703	−2.2590
CHL, HPT, SMK	−12.7198	0.0712	−2.2210
CHL, HPT	−12.7411	0.0713	−2.2613

The table shown here provides the odds ratio coefficients $\hat{\beta}$, $\hat{\delta}_1$, and $\hat{\delta}_2$ for different subsets of AGE, SMK, and ECG in the model. The first row of coefficients is for the gold standard model, which contains all five *V*s. The next row shows the coefficients obtained when SMK is dropped from the model, and so on down to the last row which shows the coefficients obtained when AGE, SMK, and ECG are simultaneously removed from the model so that only CHL and HPT are controlled.

EXAMPLE (continued)

Coefficients change somewhat. No radical change

In scanning the above table, it is seen for each coefficient separately (that is, by looking at the values in a given column) that the estimated values change somewhat as different subsets of AGE, SMK, and ECG are dropped. However, there does not appear to be a radical change in any coefficient.

Meaningful differences in $\widehat{\text{OR}}$?

- Coefficients on log odds ratio scale
- More appropriate: odds ratio scale

Nevertheless, it is not clear whether there is sufficient change in any coefficient to indicate meaningful differences in odds ratio values. Assessing the effect of a change in coefficients on odds ratio values is difficult because the coefficients are on the log odds ratio scale. It is more appropriate to make our assessment of confounding using odds ratio values rather than log odds ratio values.

$$\widehat{\text{OR}} = \exp(\hat{\beta} + \hat{\delta}_1\text{CHL} + \hat{\delta}_2\text{HPT})$$

Specify values of effect modifiers
Obtain summary table of ORs

To obtain numerical values for the odds ratio for a given model, we must specify values of the effect modifiers in the odds ratio expression. Different specifications will lead to different odds ratios. Thus, for a given model, we must consider a summary table or graph that describes the different odds ratio values that are calculated.

Compare
gold standard vs. other models
using (without *V*s)
odds ratio tables or graphs

To compare the odds ratios for two different models, say the gold standard model with the model that deletes one or more eligible *V* variables, we must compare corresponding odds ratio tables or graphs.

Evans County example:
Gold standard
vs.
Model without AGE, SMK, and ECG

As an illustration using the Evans County data, we compare odds ratio values computed from the gold standard model with values computed from the model that deletes the three eligible variables AGE, SMK, and ECG.

Gold standard $\widehat{\text{OR}}$:

$$\widehat{\text{OR}} = \exp(-12.6894 + 0.0692\text{CHL} - 2.3318\text{HPT})$$

	HTP = 0	HTP = 1
CHL = 200	$\widehat{\text{OR}} = 3.16$	$\widehat{\text{OR}} = 0.31$
CHL = 220	$\widehat{\text{OR}} = 12.61$	$\widehat{\text{OR}} = 1.22$
CHL = 240	$\widehat{\text{OR}} = 50.33$	$\widehat{\text{OR}} = 4.89$

$\text{CHL} = 200, \text{HPT} = 0 \Rightarrow \widehat{\text{OR}} = \boxed{3.16}$

$\text{CHL} = 220, \text{HPT} = 1 \Rightarrow \widehat{\text{OR}} = \boxed{1.22}$

The table shown here gives odds ratio values for the gold standard model, which contains all five *V* variables, the exposure variable CAT, and the two interaction terms CAT × CHL and CAT × HPT. In this table, we have specified three different row values for CHL, namely, 200, 220, and 240, and two column values for HPT, namely, 0 and 1. For each combination of CHL and HPT values, we thus get a different odds ratio.

EXAMPLE (continued)

$\text{CHL} = 200, \text{HPT} = 0 \Longrightarrow \widehat{\text{OR}} = 3.16$

$\text{CHL} = 220, \text{HPT} = 1 \Longrightarrow \widehat{\text{OR}} = 1.22$

For example, if CHL equals 200 and HPT equals 0, the computed odds ratio is 3.16, whereas if CHL equals 220 and HPT equals 1, the computed odds ratio is 1.22.

$\widehat{\text{OR}}$ with AGE, SMK, ECG deleted:

$$\widehat{\text{OR}}^* = \exp(-12.7411 + 0.0713\text{CHL} - 2.2613\text{HPT})$$

	HPT = 0	HPT = 1
CHL = 200	$\widehat{\text{OR}}^* = 4.57$	$\widehat{\text{OR}}^* = 0.48$
CHL = 220	$\widehat{\text{OR}}^* = 19.01$	$\widehat{\text{OR}}^* = 1.98$
CHL = 240	$\widehat{\text{OR}}^* = 79.11$	$\widehat{\text{OR}}^* = 8.34$

The table shown here gives odds ratio values, indicated by "asterisked" $\widehat{\text{OR}}$, for a model that deletes the three eligible *V* variables, AGE, SMK, and ECG. As with the gold standard model, the odds ratio expression involves the same two effect modifiers CHL and HPT, and the table shown here considers the same combination of CHL and HPT values.

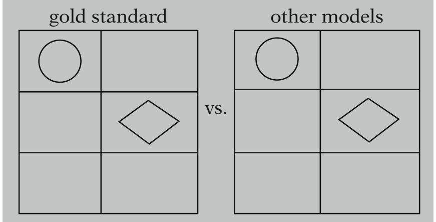
Gold standard $\widehat{\text{OR}}$: $\widehat{\text{OR}}^*$
w/o AGE, SMK, ECG

	HPT = 0	HPT = 1	HPT = 0	HPT = 1
CHL = 200	3.16	0.31	4.57	0.48
CHL = 220	12.61	1.22	19.01	1.98
CHL = 240	50.33	4.89	79.11	8.34

If we compare corresponding odds ratios in the two tables, we can see sufficient discrepancies.

For example, when CHL equals 200 and HPT equals 0, the odds ratio is 3.16 in the gold standard model, but is 4.57 when AGE, SMK, and ECG are deleted. Also, when CHL equals 220 and HPT equals 1, the corresponding odds ratios are 1.22 and 1.98.

Cannot simultaneously drop AGE, SMK, and ECG from model



Other models: delete AGE and SMK or delete AGE and ECG, etc.

Thus, because the two tables of odds ratios differ appreciably, we cannot simultaneously drop AGE, SMK, and ECG from the model.

Similar comparisons can be made by comparing the gold standard odds ratio with odds ratios obtained by deleting other subsets, for example, AGE and SMK together, or AGE and ECG together, and so on. All such comparisons show sufficient discrepancies in corresponding odds ratios. Thus, we cannot drop any of the three eligible variables from the model.

Result: cannot drop AGE, SMK, or ECG

Final model:

E: CAT
Five *V*s: CHL, HPT, AGE, SMK, ECG
Two interactions: CAT × CHL, CAT × HPT

We conclude that all five *V* variables need to be controlled, so that the final model contains the exposure variable CAT, the five *V* variables, and the interaction variables involving CHL and HPT.

EXAMPLE (continued)

No need to consider precision in this example:
Compare tables of CIs – subjective

Note that because we cannot drop either of the variables AGE, SMK, or ECG as nonconfounders, we do not need to consider possible gain in precision from deleting nonconfounders. If precision were considered, we would compare tables of confidence intervals for different models. As with confounding assessment, such comparisons are largely subjective.

Confounding and precision difficult if interaction (subjective)

Caution. Do not sacrifice validity for minor gain in precision

This example illustrates why we will find it difficult to assess confounding and precision if our model contains interaction terms. In such a case, any decision to delete possible nonconfounders is largely subjective. Therefore, we urge caution when deleting variables from our model in order to avoid sacrificing validity in exchange for what is typically only a minor gain in precision.

Summary result for final model:

Table of $\widehat{\text{OR}}$

CHL	HPT = 0	HPT = 1
200	3.16	0.31
220	12.61	1.22
240	50.33	4.89

Table of 95% CIs

CHL	HPT = 0	HPT = 1
200	(0.89, 11.03)	(0.10, 0.91)
220	(3.65, 42.94)	(0.48, 3.10)
240	(11.79, 212.23)	(1.62, 14.52)

To conclude this example, we point out that, using the final model, a summary of the results of the analysis can be made in terms of the table of odds ratios and the corresponding table of confidence intervals.

Both tables are shown here. The investigator must use this information to draw meaningful conclusions about the relationship under study. In particular, the nature of the interaction can be described in terms of the point estimates and confidence intervals.

Use to draw meaningful conclusions

$$\text{CHL}\nearrow \;\Rightarrow\; \widehat{\text{OR}}_{\text{CAT, CHD}}\nearrow$$

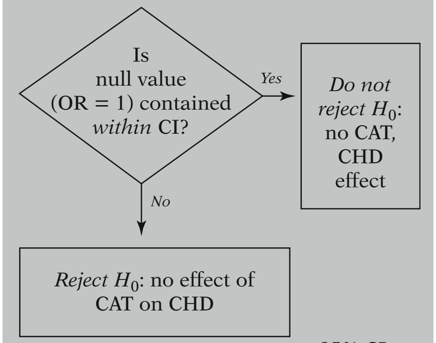
$$\text{CHL fixed: } \underset{\text{HPT = 0}}{\widehat{\text{OR}}_{\text{CAT, CHD}}} > \underset{\text{HPT = 1}}{\widehat{\text{OR}}_{\text{CAT, CHD}}}$$

All CIs are wide

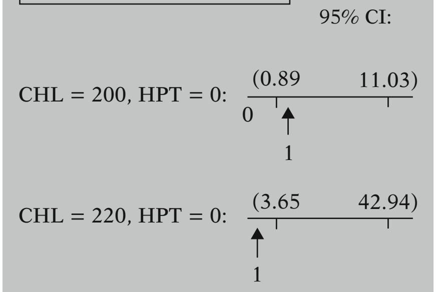
For example, as CHL increases, the odds ratio for the effect of CAT on CHD increases. Also, for fixed CHL, this odds ratio is higher when HPT is 0 than when HPT equals 1. Unfortunately, all confidence intervals are quite wide, indicating that the point estimates obtained are quite unstable.

EXAMPLE (continued)

Tests of significance:


Is
null value
(OR = 1) contained
within CI?
Yes
Do not reject H_0: no CAT, CHD effect
No
Reject H_0: no effect of CAT on CHD

Furthermore, tests of significance can be carried out using the confidence intervals. To do this, one must determine whether or not the null value of the odds ratio, namely, 1, is contained within the confidence limits. If so, we do not reject, for a given CHL, HPT combination, the null hypothesis of no effect of CAT on CHD. If the value 1 lies outside the confidence limits, we would reject the null hypothesis of no effect.



For example, when CHL equals 200 and HPT equals 0, the value of 1 is contained within the limits 0.89 and 11.03 of the 95% confidence interval. However, when CHL equals 220 and HPT equals 0, the value of 1 is not contained within the limits 3.65 and 42.94.

Test results at 5% level:

CHL = 200, HPT = 0 : no significant CAT, CHD effect

CHL = 220, HPT = 0 : significant CAT, CHD effect

Thus, when CHL equals 200 and HPT equals 0, there is no significant CAT, CHD effect, whereas when CHL equals 220 and HPT equals 0, the CAT, CHD effect is significant at the 5% level.

Tests based on CIs are *two-tailed*

In EPID, most tests of *E–D* relationship are *one-tailed*

Note that tests based on confidence intervals are two-tailed tests. One-tailed tests are more common in epidemiology for testing the effect of an exposure on disease.

One-tailed tests:

Use large sample

$$Z = \frac{\text{estimate}}{\text{standard error}}$$

When there is interaction, one-tailed tests can be obtained by using the point estimates and their standard errors that go into the computation of the confidence interval. The point estimate divided by its standard error gives a large sample Z statistic, which can be used to carry out a one-tailed test.

VI. SUMMARY

Chap. 6

- Overall guidelines for three stages
- Focus: variable specification
- HWF model

A brief summary of this presentation is now given. This has been the second of two chapters on modeling strategy when there is a single E. In Chap. 6, we gave overall guidelines for three stages, namely, variable specification, interaction assessment, and confounding assessment, with consideration of precision. Our primary focus was the variable specification stage, and an important requirement was that the initial model be hierarchically well formulated (HWF).

Chap. 7
Focus: interaction and confounding assessment
Interaction: use hierarchical backward elimination

In this chapter, we have focused on the interaction and confounding assessment stages of our modeling strategy. We have described how interaction assessment follows a hierarchical backward elimination procedure, starting with assessing higher order interaction terms followed by assessing lower order interaction terms using statistical testing methods.

Use *hierarchy principle* to identify lower order components that cannot be deleted (EVs, V_is, and V_iV_js)

If certain interaction terms are significant, we use the hierarchy principle to identify all lower order components of such terms, which cannot be deleted from any further model considered. This applies to lower order interaction terms (i.e., terms of the form EV) and to lower order terms involving potential confounders of the form V_i or V_iV_j.

Confounding: *no* statistical testing: Compare whether $\widehat{\text{OR}}$ meaningfully changes when Vs are deleted

Confounding is assessed without the use of statistical testing. The procedure involves determining whether the estimated odds ratio meaningfully changes when eligible V variables are deleted from the model.

Drop nonconfounders if precision is gained by examining CIs

If some variables can be identified as nonconfounders, they may be dropped from the model provided their deletion leads to a gain in precision from examining confidence intervals.

No interaction: assess confounding by monitoring changes in $\hat{\beta}$, the coefficient of E

If there is no interaction, the assessment of confounding is carried out by monitoring changes in the estimated coefficient of the exposure variable.

SUMMARY (*continued*)

Interaction present: compare tables of odds ratios and confidence intervals (subjective)

However, if there is interaction, the assessment of confounding is much more subjective because it typically requires the comparison of tables of odds ratio values. Similarly, assessing precision requires comparison of tables of confidence intervals.

Interaction: Safe (for validity) to keep all *V*s in model

Consequently, if there is interaction, it is typically safe for ensuring validity to keep all potential confounders in the model, even those that are candidates to be deleted as possible nonconfounders.

Chapters

1. Introduction
2. Special Cases

•

•

✓ 7. Interaction and Confounding Assessment

8. Additional Modeling Strategy Issues

This presentation is now complete. The reader may wish to review the detailed summary and to try the practice exercises and test that follow.

The next chapter considers additional issues about modeling strategy, including how to address more than one exposure variable, screening variables, collinearity, multiple testing, and influential observations.

Detailed Outline

I. **Overview** (pages 206–207)

Focus:

- Assessing confounding and interaction
- Obtaining a valid estimate of the E–D relationship

A. Three stages: variable specification, interaction assessment, and confounding assessment followed by consideration of precision.

B. Variable specification stage

  i. Start with D, E, and $C_1, C_2, \ldots, C_p$.

  ii. Choose Vs from Cs based on prior research or theory and considering potential statistical problems, e.g., collinearity; simplest choice is to let Vs be Cs themselves.

  iii. Choose Ws from Cs to be either Vs or product of two Vs; usually recommend Ws to be Cs themselves or some subset of Cs.

C. The model must be *hierarchically well formulated* (HWF): given any variable in the model, all lower order components must also be in the model.

D. The strategy is a *hierarchical backward elimination strategy*: evaluate EV_iV_j terms first, then V_i terms, then V_i terms last.

E. The *hierarchy principle* needs to be applied for any variable kept in the model: If a variable is to be retained in the model, then all lower order components of that variable are to be retained in all further models considered.

II. **Interaction assessment stage** (pages 207–210)

A. Flow diagram representation.

B. Description of flow diagram: test higher order interactions first, then apply hierarchy principle, then test lower order interactions.

C. How to carry out tests: chunk tests first, followed by backward elimination whether or not chunk test is significant; testing procedure involves likelihood ratio statistic.

D. Example.

III. **Confounding and precision assessment when no interaction** (pages 211–215)

A. Monitor changes in the effect measure (the odds ratio) corresponding to dropping subsets of potential confounders from the model.

B. Gold standard odds ratio obtained from model containing all Vs specified initially.

C. Identify subsets of Vs giving approximately the same odds ratio as gold standard.

D. Control for the subset that gives largest gain in precision, i.e., tighter confidence interval around odds ratio.

E. Example.

IV. Confounding assessment with interaction (pages 215–223)

A. Flow diagram representation.

B. Use hierarchy principle to identify all *V*s that cannot be eliminated from the model; the remaining *V*s are eligible to be dropped.

C. Eligible *V*s can be dropped as nonconfounders if odds ratio does not change when dropped; then control for subset of remaining *V*s that gives largest gain in precision.

D. Alternative ways to determine whether odds ratio changes when different subsets of *V*s are dropped.

E. In practice, it is difficult to evaluate changes in odds ratio when eligible *V*s are dropped; consequently, safest strategy is to control for all *V*s.

F. Example.

V. Evans County example continued (pages 223–230)

A. Evans County CHD data descriptions.

B. Variable specification stage.

C. Confounding assessment stage.

D. Final model results.

Practice Exercises

A prevalence study of predictors of surgical wound infection in 265 hospitals throughout Australia collected data on 12,742 surgical patients (McLaws et al., 1988). For each patient, the following independent variables were determined: type of hospital (public or private), size of hospital (large or small), degree of contamination of surgical site (clean or contaminated), and age and sex of the patient. A logistic model was fit to this data to predict whether or not the patient developed a surgical wound infection during hospitalization. The abbreviated variable names and the manner in which the variables were coded in the model are described as follows:

Variable	Abbreviation	Coding
Type of hospital	HT	1 = public, 0 = private
Size of hospital	HS	1 = large, 0 = small
Degree of contamination	CT	1 = contaminated, 0 = clean
Age	AGE	Continuous
Sex	SEX	1 = female, 0 = male

1. Suppose the following initial model is specified for assessing the effect of type of hospital (HT), considered as the exposure variable, on the prevalence of surgical wound infection, controlling for the other four variables on the above list:

   $$\begin{aligned}\text{logit P}(\mathbf{X}) = {} & \alpha + \beta\text{HT} + \gamma_1\text{HS} + \gamma_2\text{CT} + \gamma_3\text{AGE} + \gamma_4\text{SEX} \\ & + \delta_1\text{HT} \times \text{AGE} + \delta_2\text{HT} \times \text{SEX}.\end{aligned}$$

   Describe how to test for the overall significance (a "chunk" test) of the interaction terms. In answering this, describe the null hypothesis, the full and reduced models, the form of the test statistic, and its distribution under the null hypothesis.
2. Using the model given in Exercise 1, describe briefly how to carry out a backward elimination procedure to assess interaction.
3. Briefly describe how to carry out interaction assessment for the model described in Exercise 1. (In answering this, it is suggested you make use of the tests described in Exercises 1 and 2.)
4. Suppose the interaction assessment stage for the model in Example 1 finds no significant interaction terms. What is the formula for the odds ratio for the effect of HT on the prevalence of surgical wound infection at the end of the interaction assessment stage? What *V* terms remain in the model at the end of interaction assessment? Describe how you would evaluate which of these *V* terms should be controlled as confounders.
5. Considering the scenario described in Exercise 4 (i.e., no interaction terms found significant), suppose you determine that the variables CT and AGE do not need to be controlled for confounding. Describe how you would consider whether dropping both variables will improve precision.
6. Suppose the interaction assessment stage finds that the interaction terms HT × AGE and HT × SEX are both significant. Based on this result, what is the formula for the odds ratio that describes the effect of HT on the prevalence of surgical wound infection?
7. For the scenario described in Example 6, and making use of the hierarchy principle, what *V* terms are eligible to be dropped as possible nonconfounders?
8. Describe briefly how you would assess confounding for the model considered in Exercises 6 and 7.
9. Suppose that the variable CT is determined to be a *non*confounder, whereas all other *V* variables in the model (of Exercise 1) need to be controlled. Describe

briefly how you would assess whether the variable CT needs to be controlled for precision reasons.

10. What problems are associated with the assessment of confounding and precision described in Exercises 8 and 9?

Test

The following questions consider the use of logistic regression on data obtained from a matched case-control study of cervical cancer in 313 women from Sydney, Australia (Brock et al., 1988). The outcome variable is cervical cancer status (1 = present, 0 = absent). The matching variables are age and socioeconomic status. Additional independent variables not matched on are smoking status, number of lifetime sexual partners, and age at first sexual intercourse. The independent variables are listed below together with their computer abbreviation and coding scheme.

Variable	Abbreviation	Coding
Smoking status	SMK	1 = ever, 0 = never
Number of sexual partners	NS	1 = 4+, 0 = 0–3
Age at first intercourse	AS	1 = 20+, 0 = ≤19
Age of subject	AGE	Category matched
Socioeconomic status	SES	Category matched

Assume that at the end of the variable specification stage, the following *E*, *V*, *W* model has been defined as the initial model to be considered:

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha + \beta\text{SMK} + \textstyle\sum \gamma_i^* V_i^* + \gamma_1\text{NS} + \gamma_2\text{AS} \\ + \gamma_3\text{NS} \times \text{AS} + \delta_1\text{SMK} \times \text{NS} + \delta_2\text{SMK} \times \text{AS} \\ + \delta_3\text{SMK} \times \text{NS} \times \text{AS},\end{aligned}$$

where the V_i^* are dummy variables indicating matching strata, the γ_i^* are the coefficients of the V_i^* variables, SMK is the only exposure variable of interest, and the variables NS, AS, AGE, and SES are being considered for control.

1. For the above model, which variables are interaction terms?
2. For the above model, list the steps you would take to assess interaction using a hierarchically backward elimination approach.
3. Assume that at the end of interaction assessment, the only interaction term found significant is the product term SMK × NS. What variables are left in the model at

the end of the interaction stage? Which of the *V* variables in the model cannot be deleted from any further models considered? Explain briefly your answer to the latter question.

4. Based on the scenario described in Question 3 (i.e., the only significant interaction term is SMK × NS), what is the expression for the odds ratio that describes the effect of SMK on cervical cancer status at the end of the interaction assessment stage?
5. Based again on the scenario described in Question 3, what is the expression for the odds ratio that describes the effect of SMK on cervical cancer status if the variable NS × AS is dropped from the model that remains at the end of the interaction assessment stage?
6. Based again on the scenario described in Question 3, how would you assess whether the variable NS × AS should be retained in the model? (In answering this question, consider both confounding and precision issues.)
7. Suppose the variable NS × AS is dropped from the model based on the scenario described in Question 3. Describe how you would assess confounding and precision for any other V terms still eligible to be deleted from the model after interaction assessment.
8. Suppose the final model obtained from the cervical cancer study data is given by the following printout results:

Variable	β	S.E.	Chi sq	*P*
SMK	1.9381	0.4312	20.20	0.0000
NS	1.4963	0.4372	11.71	0.0006
AS	−0.6811	0.3473	3.85	0.0499
SMK × NS	−1.1128	0.5997	3.44	0.0635

Describe briefly how you would use the above information to summarize the results of your study. (In your answer, you need only describe the information to be used rather than actually calculate numerical results.)

Answers to Practice Exercises

1. A "chunk" test for overall significance of interaction terms can be carried out using a likelihood ratio test that compares the initial (full) model with a reduced model under the null hypothesis of no interaction terms. The likelihood ratio test will be a chi-square test with two degrees of freedom (because two interaction terms are being tested simultaneously).

2. Using a backward elimination procedure, one first determines which of the two product terms HT × AGE and HT × SEX is the least significant in a model containing these terms and all main effect terms. If this least significant term is significant, then both interaction terms are retained in the model. If the least significant term is nonsignificant, it is then dropped from the model. The model is then refitted with the remaining product term and all main effects. In the refitted model, the remaining interaction term is tested for significance. If significant, it is retained; if not significant, it is dropped.
3. Interaction assessment would be carried out first using a "chunk" test for overall interaction as described in Exercise 1. If this test is not significant, one could drop both interaction terms from the model as being not significant overall. If the chunk test is significant, then backward elimination, as described in Exercise 2, can be carried out to decide if both interaction terms need to be retained or whether one of the terms can be dropped. Also, even if the chunk test is not significant, backward elimination may be carried out to determine whether a significant interaction term can still be found despite the chunk test results.
4. The odds ratio formula is given by exp(β), where β is the coefficient of the HT variable. All V variables remain in the model at the end of the interaction assessment stage. These are HS, CT, AGE, and SEX. To evaluate which of these terms are confounders, one has to consider whether the odds ratio given by exp(β) changes as one or more of the V variables are dropped from the model. If, for example, HS and CT are dropped and exp(β) does not change from the (gold standard) model containing all Vs, then HS and CT do not need to be controlled as confounders. Ideally, one should consider as candidates for control any subset of the four V variables that will give the same odds ratio as the gold standard.
5. If CT and AGE do not need to be controlled for confounding, then, to assess precision, we must look at the confidence intervals around the odds ratio for a model which contains neither CT nor AGE. If this confidence interval is meaningfully narrower than the corresponding confidence interval around the gold standard odds ratio, then precision is gained by dropping CT and AGE. Otherwise, even though these variables need not be controlled for confounding, they

should be retained in the model if precision is not gained by dropping them.

6. The odds ratio formula is given by exp(β + δ_1AGE + δ_2SEX).
7. Using the hierarchy principle, CT and HS are eligible to be dropped as nonconfounders.
8. Drop CT, HS, or both CT and HS from the model and determine whether the coefficients β, δ_1, and δ_2 in the odds ratio expression change. Alternatively, determine whether the odds ratio itself changes by comparing tables of odds ratios for specified values of the effect modifiers AGE and SEX. If there is no change in coefficients and/or in odds ratio tables, then the variables dropped do not need to be controlled for confounding.
9. Drop CT from the model and determine if the confidence interval around the odds ratio is wider than the corresponding confidence interval for the model that contains CT. Because the odds ratio is defined by the expression exp(β + δ_1AGE + δ_2SEX), a table of confidence intervals for both the model without CT and with CT will need to be obtained by specifying different values for the effect modifiers AGE and SEX. To assess whether CT needs to be controlled for precision reasons, one must compare these tables of confidence intervals. If the confidence intervals when CT is not in the model are narrower in some overall sense than when CT is in the model, precision is gained by dropping CT. Otherwise, CT should be controlled as precision is not gained when the CT variable is removed.
10. Assessing confounding and precision in Exercises 8 and 9 requires subjective comparisons of either several regression coefficients, several odds ratios, or several confidence intervals. Such subjective comparisons are likely to lead to highly debatable conclusions, so that a safe course of action is to control for all *V* variables regardless of whether they are confounders or not.

8 Additional Modeling Strategy Issues

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_8, © Springer Science+Business Media, LLC 2010

Introduction

In this chapter, we consider five issues on modeling Strategy, which were not covered in the previous two chapters on this topic:

1. Modeling strategy when there are two or more exposure variables
2. Screening variables when modeling
3. Collinearity diagnostics
4. Influential observations
5. Multiple testing

Each of these issues represent important features of any regression analysis that typically require attention when determining a "best" model, although our specific focus concerns a binary logistic regression model.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

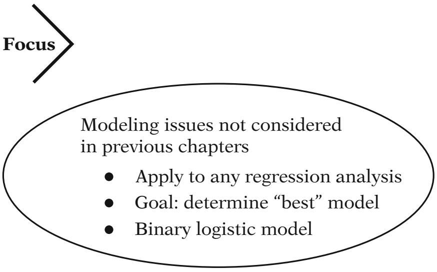
Objectives

Upon completing this chapter, the learner should be able to:

1. Given a binary logistic model involving two or more exposures, describe or illustrate how to carry out a modeling strategy to determine a "best" model.
2. Given a fitted binary logistic model involving a large number of exposure and/or covariates (potential confounders or effect modifiers), describe or illustrate how to conduct screening to reduce the number of variables to be considered in your initial multivariate model.
3. Explain by illustration when it is questionable to screen covariates using statistical testing for a crude association with the outcome variable.
4. Given a binary logistic model involving several exposures and/or covariates, describe and/or illustrate how to assess collinearity and how to proceed if a collinearity problem is identified.
5. Given a binary logistic model involving several exposure variables and/or covariates, describe and/or illustrate how to determine whether there are any influential observations and how to proceed with the analysis if influential observations are found.
6. Given a binary logistic model involving several exposure variables and/or covariates, describe and/or illustrate how to consider (or possibly correct for) multiple testing when carrying out a modeling strategy to determine a "best" model.

Presentation

I. Overview



Modeling issues not considered in previous chapters

- Apply to any regression analysis
- Goal: determine "best" model
- Binary logistic model

This presentation addresses several modeling strategy issues not considered in the previous two chapters (6 and 7). These issues represent important features of *any* regression analysis that typically require attention when going about the process of determining a "best" model, although our specific focus concerns a *binary logistic regression model*.

Issues:

1. Modeling strategy when there are two or more exposure variables
2. Screening variables when modeling
3. Collinearity diagnostics
4. Influential observations
5. Multiple testing

We consider five issues, listed here at the left, each of which will be described and illustrated in the sections that follow.

II. Modeling Strategy for Several Exposure Variables

Extend modeling strategy

- Outome: $D(0,1)$
- Exposures: $E_1, E_2, \ldots, E_\text{q}$
- Control variables: $C_1, C_2, \ldots, C_\text{p}$

In this section, we extend the modeling strategy guidelines described in the previous two chapters to consider two or more exposure variables, controlling for covariates that are potential confounders and/or effect modifiers. We begin with an example involving exactly two exposure variables.

EXAMPLE

Example: Two *E*s

Cross-sectional study
Grady Hospital, Atlanta, GA
297 adult patients
Diagnosis: *Staph. aureus* infection

Concern: potential predictors of MRSA

A cross-sectional study carried out at Grady Hospital in Atlanta, Georgia involved 297 adult patients seen in an emergency department whose blood cultures taken within 24 hours of admission were found to have *Staphylococcus aureus* infection (Rezende et al., 2002). Information was obtained on several variables that were considered as potential predictors of methicillin-resistance infection (MRSA).

EXAMPLE (continued)

Outcome: D = MRSA status
(0 = no, 1 = yes)
Predictors:

PREVHOSP (0 = no, 1 = yes)
PAMU (0 = no, 1 = yes)
AGE (continuous)
GENDER (0 = F, 1 = M)

Question:

PREVHOSP, PAMU $\Longrightarrow$ MRSA
controlling for AGE, GENDER

Two Es : E_1 = PREVHOSP
E_2 = PAMU
Two Cs : C_1 = AGE
C_2 = GENDER

Initial model:

$$\begin{aligned}\text{Logit P}(\mathbf{X}) = \alpha &+ (\beta_1 E_1 + \beta_2 E_2)\\ &+ (\gamma_1 V_1 + \gamma_2 V_2)\\ &+ (\delta_{11} E_1 W_1 + \delta_{12} E_1 W_2\\ &+ \delta_{21} E_2 W_1 + \delta_{22} E_2 W_2)\\ &+ \delta^* E_1 E_2,\end{aligned}$$

where $V_1 = C_1 = W_1$ and
$V_2 = C_2 = W_2$.

V_1 and V_2: potential confounders
W_1 and W_2: potential effect modifiers
E_1E_2: interaction of exposures

The outcome variable is MRSA status (1 = yes, 0 = no), and covariates of interest included the following variables: PREVHOSP (1 = previous hospitalization, 0 = no previous hospitalization), PAMU (1 = antimicrobial drug use in the previous 3 months, 0 = no previous antimicrobial drug use), AGE (continuous), and GENDER (1 = male, 0 = female).

For these data, we consider the following question: Are the variables PREVHOSP and PAMU associated with MRSA outcome controlling for AGE and GENDER?

For this question, our predictors include two Es (PREVHOSP and PAMU) and two Cs (AGE and GENDER).

We now consider an initial EVW model (shown at the left) that includes both Es and both Cs as main effects plus product terms involving each E with each C and the product of the two Es.

This initial model considers the control variables AGE and GENDER as both potential confounders (i.e., V_1 and V_2) and as potential effect modifiers (i.e., W_1 and W_2) of both E_1 and E_2. The model also contains an interaction term involving the two Es.

Modeling Strategy with Several Exposures

Step 1: Variable Specification (Initial Model)

Considers the following:

- Study question
- Literature review
- Biological/medical conceptualization

(previously recommended with only one E)

As recommended in the previous chapters when only one exposure variable was being considered, we continue to emphasize that the first step in one's modeling strategy, even with two or more Es, is to specify the initial model. This step requires consideration of the literature about the study question and/or outcome and/or variables needing to be controlled based on one's biological/medical conceptualization of the study question.

EXAMPLE

Our example assumes:

- AGE and GENDER risk factors
- AGE and GENDER potential effect modifiers of interest
- Interaction of PREVHOSP and PAMU also of interest

For our example, therefore, we have assumed that AGE and GENDER are well-known risk factors for MRSA, and that there is also interest to assess whether each of these variables are effect modifiers of either or both of the exposure variables. We also assume that the interaction of the exposures with each other is of interest.

No interaction model:

$$\text{Logit } P(\mathbf{X}) = \alpha + (\beta_1 E_1 + \beta_2 E_2) + (\gamma_1 V_1 + \gamma_2 V_2)$$

If, on the other hand, we decided that interaction of any kind was either not of interest or not practically interpretable, our initial model would omit such interaction terms. In such a case, the initial model would still involve the two *E*s, but it would be a no interaction model, as shown at the left.

Distinction between one *E* and several *E*s:

Single *E*: Only one type of interaction, *EW*s

Several *E*s:

Two types of interaction, E_iW_js and E_iE_ks

Potentially omit *E*s from final model

The primary distinction between the modeling strategy for a single *E* variable vs. several *E*s is that, in the latter situation, there are two types of interactions to consider: interactions of *E*s with *W*s and interactions of *E*s with other *E*s. Also, when there are several *E*s, we may consider omitting some *E*s (as nonsignificant) in the final ("best") model.

General form: EVW model for several *E*s

$$\text{Logit } P(\mathbf{X}) = \alpha + \sum_{i=1}^{q} \beta_i E_i + \sum_{j=1}^{p_1} \gamma_j V_j + \sum_{i=1}^{q} \sum_{k=1}^{p_2} \delta_{ik} E_i W_k + \sum_{i=1}^{q} \sum_{\substack{i'=1 \\ i \neq i'}}^{q} \delta^*_{ii'} E_i E_{i'}$$

Although we will return to this example shortly, we show here at the left the general form of the EVW model when there are several exposures. This model is written rather succinctly using summation signs, including double summation signs when considering interactions. Notice that in this general form, there are q exposure variables, p_1 potential confounders, and p_2 potential effect modifiers.

Alternative form (w/o summation signs):

This same model is alternatively written without summation signs here and is divided into four groups of predictor variables:

$$\text{Logit P(X)} = \boxed{\alpha + \beta_1 E_1 + \beta_2 E_2 + \ldots + \beta_q E_q} \quad \textbf{\textit{E}s}$$

The first group lists the *E* variables.

$$\boxed{+ \gamma_1 V_1 + \gamma_2 V_2 + \ldots + \gamma_{p_1} V_{p_1}} \quad \textbf{\textit{V}s}$$

The second group lists the *V* variables.

$$\boxed{\begin{aligned} &+ \delta_{11} E_1 W_1 + \delta_{12} E_1 W_2 + \ldots + \delta_{1,p_2} E_1 W_{p_2} \\ &+ \delta_{21} E_2 W_1 + \delta_{22} E_2 W_2 + \ldots + \delta_{2,p_2} E_2 W_{p_2} \\ &+ \ldots \\ &+ \delta_{q1} E_q W_1 + \delta_{q2} E_q W_2 + \ldots + \delta_{q,p_2} E_q W_{p_2} \end{aligned}} \quad \textbf{\textit{EW}s}$$

The third group lists the *EW* variables, the first line of which contains products of E_1 with each of the W_js, the second line contains products of E_2 with each of the W_js, and so on, with the last line of the group containing products of E_q with each of the W_js.

$$\boxed{\begin{aligned} &+ \delta^*_{12} E_1 E_2 + \delta^*_{13} E_1 E_3 + \ldots + \delta^*_{1q} E_1 E_q \\ &+ \delta^*_{23} E_2 E_3 + \delta^*_{24} E_2 E_4 + \ldots + \delta^*_{2q} E_2 E_q \\ &+ \ldots + \delta^*_{q-1,q} E_{q-1} E_q \end{aligned}} \quad \begin{matrix}\textbf{\textit{EE}s} \\ \boldsymbol{i \neq i'}\end{matrix}$$

Finally, the fourth group lists the *EE* variables, with the first line containing products of E_1 with all other *E*s, the second line containing products of E_2 with all other *E*s except E_1, and so on, with the last line containing the single product term $E_{q-1}\ E_q$.

EXAMPLE

MRSA example:

$$\text{Logit P(X)} = \alpha + \beta_1 E_1 + \beta_2 E_2 \quad \boxed{q = 2}$$

$$+ \gamma_1 V_1 + \gamma_2 V_2 \quad \boxed{p_1 = 2}$$

$$+ \delta_{11} E_1 W_1 + \delta_{12} E_1 W_2 + \delta_{21} E_2 W_1$$

$$+ \delta_{22} E_2 W_2 \quad \boxed{4\ EWs}$$

$$+ \delta^* E_1 E_2 \quad \boxed{1\ EE}$$

Returning to our initial model for the MRSA data, there are $q = 2$ *E* variables,

$p_1 = 2$ *V* variables,

$p_2 = 2$ *W* variables, which yields 4 *EW* variables, and a single *EE* variable.

Next step in modeling strategy?

Step 2: Assess interaction

Questions regarding *EW*s and *EE*s?

- Consider separately or simultaneously?
- If separately, *EW*s or *EE*s first?

So, how do we proceed once we have identified our initial model? Following our previous strategy for one *E* variable, we recommend assessing interaction as the next step. But, since there are two types of product terms, *EW*s and *EE*s, should we consider these types separately or simultaneously, and if separately, do we first consider *EW*s or *EE*s?

Answer: It depends!
Several reasonable options.

The answer, not surprisingly, is it depends! That is, there are several reasonable options.

Option A: Overall (chunk) LR test for interaction; then "subchunk" LR tests for *EW*s and *EE*s; then *V*s; finally *E*s

One Option (**A**) begins with a "chunk" LR test that simultaneously evaluates all product terms. We then test separate "subchunks" involving *EW*s and *EE*s, after which we assess the *V*s for confounding and precision. Finally, we consider dropping nonsignificant *E*s.

EXAMPLE

MRSA example:

- **Overall chunk test:** LR $\sim \chi^2{}_{5\ \text{df}}$ under H_0: $\delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = \delta^* = 0$
- **Subchunk tests:** LR $\sim \chi^2{}_{4\ \text{df}}$ under H_{01}: $\delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = 0$
  LR $\sim \chi^2{}_{1\ \text{df}}$ under H_{02}: $\delta^* = 0$

For the initial MRSA model, since there are five product terms, the overall chunk test would involve a chi square statistic with 5 degrees of freedom. The two "subchunks" would involve the 4 *EW* terms and the single *EE* term, as we illustrate on the left.

Option B: Assess *EW*s first, then *EE*s, prior to *V*s and *E*s

Reasons:
Assess interaction (*EW*s and *EE*s) prior to confounding and precision, and *Assess EW*s prior to *EE*s

Alternatively, a second Option (**B**) differs from Option **A** by simply skipping the overall chunk test. Both Options A and B make sense if we decide that assessing interaction should always precede assessing confounding and precision, and that *EW*s should always be assessed prior to *EE*s.

Option C: Assess *EW*s first, then *V*s, prior to *EE*s and *E*s

Reason:
Assess effect modification (*W*s) and confounding (*V*s) before considering exposures (*E*s and *EE*s)

As another Option (**C**), recall that when we considered a model with only a single *E*, we left this *E* in the model throughout the entire process of evaluating interaction, confounding, and then precision. An analogous approach for several *E*s is to evaluate effect modifiers (*W*s) and potential confounders (*V*s) before considering any terms involving *E*s, including product terms (*EE*s).

EXAMPLE

Initial Model Output:
$-2\ln L = 275.683$

```
Analysis of maximum likelihood estimates
        Param      DF  Estimate  StdErr  ChiSq   Pr > ChiSq
        Intercept  1   -3.8214   1.3594  7.9018  0.0049
  Es   {PREVHOSP   1    1.0027   1.9063  0.2767  0.5989
       {PAMU       1   -0.6216   1.7797  0.1220  0.7269
  Vs   {AGE        1    0.0240   0.0186  1.6626  0.1973
       {GENDER     1    0.3968   0.7727  0.2638  0.6075
       {PRHAGE     1    0.00471  0.0249  0.0359  0.8498
  EWs  {PRHGEN     1   -0.2533   1.058   0.0573  0.8108
       {PAMAGE     1    0.0124   0.0227  0.2971  0.5857
       {PAMGEN     1    1.1335   0.9138  1.5384  0.2149
  EEs  {PRHPAM     1    1.2065   0.9297  1.6841  0.1944
```

We will apply Options **A** through **C** to the MRSA data. First, we present, at the left, edited results from fitting the initial model.

Reduced Model A:

$$\text{Logit P}(\mathbf{X}) = \alpha + (\beta_1 E_1 + \beta_2 E_2) + (\gamma_1 V_1 + \gamma_2 V_2)$$

We now show the results from using *Option* **A** for the reduced model (**A**) that eliminates all five interaction terms from the initial model.

EXAMPLE

Model A Output: $-2\ln L = 279.317$

```
Analysis of maximum likelihood estimates
       Param      DF  Estimate  StdErr   ChiSq    Pr > ChiSq
       Intercept  1   -5.0583   0.7643  43.8059  <.0001
Es  {  PREVHOSP   1    1.4855   0.4032  13.5745  0.0002
       PAMU       1    1.7819   0.3707  23.1113  <.0001
Vs  {  AGE        1    0.0353   0.0092  14.7004  0.0001
       GENDER     1    0.9329   0.3418   7.4513  0.0063
```

$$\begin{aligned} \text{LR} &= -2\ 1\text{n}\ \text{L}_{\text{R(A)}} - (-2\ 1\text{n}\ \text{L}_{\text{F}}) \\ &= 279.317 - 275.683 \\ &= \mathbf{3.634}_{\ 5\ \text{df}}\ (P = 0.6032) \end{aligned}$$

$\Rightarrow$ No–interaction model A preferred to full interaction model

The LR statistic for the overall "chunk" test that compares initial and reduced models yields a chisquare statistic of 3.634 with 5 df, which is highly nonsignificant. This suggests that the *no-interaction MRSA model A is preferable* to the initial model containing five interaction terms.

Possibility: some product terms significant

⇓

Carry out subchunk tests for *EW*s and *EE*s (start of **Option B**)

Nevertheless, to consider the possibility that some of the product terms are significant despite the nonsignificant overall chunk test results, we now carry out a test for the "subchunk" of *EW*s terms followed by another test for the "subchunk" of *EE* terms (of which there is only one: E_1E_2). We are now essentially considering (the start of) *Option* **B**.

Reduced Model B (w/o *EW* terms):

$$\begin{aligned} \text{Logit P}(\mathbf{X}) = \alpha &+ (\beta_1 E_1 + \beta_2 E_2) \\ &+ (\gamma_1 V_1 + \gamma_2 V_2) \\ &+ \delta^* E_1 E_2 \end{aligned}$$

Testing for the *EW terms* first, we present, at the left, the reduced model (**B**) obtained by eliminating the four *EW* terms from the initial model, thereby keeping the single *EE* term in the model.

Model B Output: $-2\ln L = 277.667$

```
Analysis of maximum likelihood estimates
       Param      DF  Estimate  StdErr   ChiSq    Pr > ChiSq
       Intercept  1   -4.9038   0.7649  41.1020  <.0001
Es  {  PREVHOSP   1    1.0503   0.5257   3.9922  0.0457
       PAMU       1   -0.9772   0.7589   1.6583  0.1978
Vs  {  AGE        1    0.0357   0.0092  15.0206  0.0001
       GENDER     1    0.9778   0.3464   7.9660  0.0048
EE  {  PRHPAM     1    1.0894   0.8733   1.5562  0.2122
```

The resulting output for this model is shown here.

$$\begin{aligned} \text{LR} &= -2\ln \text{L}_{\text{R(B)}} - (-2\ln \text{L}_{\text{F}}) \\ &= 277.667 - 275.683 \\ &= 1.984_{\ 4\ \text{df}}\ (P = 0.7387) \end{aligned}$$

$\Rightarrow$ Reduced model B preferred to full interaction model

From the output, the LR statistic for the "subchunk" test that compares the initial model with reduced model **B** yields a chi-square statistic of 1.984 with 4 df, which is highly nonsignificant. This suggests that the *reduced model B* (*that excludes all EW terms*) *is preferable* to the initial modeling containing five interaction terms.

EE term: E_1E_2 (=PRHPAM)
Testing H_0: $\delta^* = 0$ in reduced model **B**

Wald statistic $= 1.089_{\text{df}=1}$ ($P = 0.2122$)

LR statistic $= 279.317 - 277.667 = 1.650_{\text{df}=1}$ ($P = 0.1990$)

Focusing on *the single EE term* (E_1E_2) in the reduced model **B**, we can see from the output for this model that the Wald test for H_0: $\delta^* = 0$ is nonsignificant ($P = 0.2122$). The corresponding LR statistic (that compares model **A** with model **B**) is also nonsignificant ($P = 0.1990$).

No-interaction Model **A:**

$$\begin{aligned} \text{Logit } P(\mathbf{X}) = \alpha &+ (\beta_1 E_1 + \beta_2 E_2) \\ &+ (\gamma_1 V_1 + \gamma_2 V_2) \end{aligned}$$

preferred to interaction model

The above interaction results using *Options* **A** *and* **B** indicate that the no-interaction model **A** shown at the left is preferable to a model involving any *EW* or *EE* terms.

EXAMPLE (continued)

Options A and B (continued)

Confounding:
Does $\hat{OR}$ meaningfully change when AGE and/or GENDER are dropped?

GS model: no-interaction model A above

$$OR_{GS(A)} = \exp[\beta_1(E_1{}^* - E_1) + \beta_2(E_2{}^* - E_2)],$$

where $\mathbf{X}^* = (E_1{}^*,\ E_2{}^*)$ and $\mathbf{X} = (E_1,\ E_2)$ are two specifications of the two Es

Our choices for E_1 and E_2 on two subjects:

$$\mathbf{X}^* = (E_1{}^* = \underset{\text{yes}}{1}, E_2{}^* = \underset{\text{yes}}{1})\ \text{vs.}$$

$$\mathbf{X} = (E_1 = \underset{\text{no}}{0}, E_2 = \underset{\text{no}}{0})$$

$$\begin{aligned} OR_{GS(A)} &= \exp[\beta_1(1-0) + \beta_2(1-0)] \\ &= \exp[\beta_1 + \beta_2] \end{aligned}$$

Table of $\widehat{\mathbf{OR}}$s (check confounding)

Vs in model	AGE,GEN	AGE	GEN	Neither
$\widehat{OR}$	$\widehat{OR}_I$	$\widehat{OR}_{II}$	$\widehat{OR}_{III}$	$\widehat{OR}_{IV}$

Model #

$$\text{I. Logit } P_I(\mathbf{X}) = \alpha + \beta_1 E_1 + \beta_2 E_2 + \gamma_1 V_1 + \gamma_2 V_2$$

$$\text{II. Logit } P_{II}(\mathbf{X}) = \alpha + \beta_1 E_1 + \beta_2 E_2 + \gamma_1 V_1$$

$$\text{III. Logit } P_{III}(\mathbf{X}) = \alpha + \beta_1 E_1 + \beta_2 E_2 + \gamma_2 V_2$$

$$\text{IV. Logit } P_{IV}(\mathbf{X}) = \alpha + \beta_1 E_1 + \beta_2 E_2$$

OR formula ($E_1{}^* = 1$, $E_2{}^* = 1$) vs. ($E_1 = 0$, $E_2 = 0$) for all four models:

$$\boxed{OR = \exp[\beta_1 + \beta_2]}$$

To assess confounding, we need to determine whether the estimated OR meaningfully changes (e.g., by more than 10%) when either AGE or GENDER or both are dropped from the model. Here, the gold standard (GS) model is the no-interaction model **A** just shown.

The formula for the odds ratio for the GS model is shown at the left, where (E_1*, E_2*) and (E_1, E_2) denote two specifications of the two exposures PREVHOSP (i.e., E_1) and PAMU (i.e., E_2).

There are several ways to specify **X*** and **X** for PREVHOSP and PAMU. Here, for convenience and simplicity, we will choose to compare a subject **X*** who is positive (i.e., yes) for both Es with a subject **X** who is negative (i.e., no) for both Es.

Based on the above choices, the OR formula for our GS reduced model **A** simplifies, as shown here.

To assess confounding, we must now determine whether estimates of our simplified $OR_{GS(\mathbf{A})}$ meaningfully change when we drop AGE and/or GENDER. This requires us to consider a table of ORs, as shown at the left.

To complete the above table, we need to fit the four models shown at the left. The first model, which we have already described, is the GS(**A**) model containing PREVHOSP, PAMU, AGE, and GENDER. The other three models exclude GENDER, AGE, or both from the model.

Since all four models involve the same two E variables, the general formula for the OR that compares a subject who is exposed on both Es ($E_1{}^* = 1$, $E_2{}^* = 1$) vs. a subject who is not exposed on both Es ($E_1 = 0$, $E_2 = 0$) has the same algebraic form for each model, including the GS model.

EXAMPLE (continued)

Options A and B (continued)

However: $\hat{\beta}_1$ and $\hat{\beta}_2$ likely differ for each model

However, since the models do not all have the same predictors, the estimates of the regression coefficients are likely to differ somewhat.

Estimate Regression Coefficients and $\widehat{\text{OR}}$s

Model:	I (GS)	II	III	IV
*V*s in model	AGE,GEN	AGE	GEN	Neither
$\hat{\beta}_1$	1.4855	1.4679	1.6627	1.6363
$\hat{\beta}_2$	1.7819	1.6394	1.4973	1.4542
$\widehat{\text{OR}}$	26.2430	22.3606	23.5706	21.9881

At the left, we show for each model, the values of these two estimated regression coefficients together with their corresponding OR estimates. *From this information, we must decide which one or more of the four models controls for confounding*. Certainly, the GS model controls for confounding, but do any of the other models do so also?

Assessing confounding (Option **B**): Which models have "same" $\widehat{\text{OR}}$ as GS model?

Quick glance: OR for GS highest

⇓

Only GS model controls confounding

An equivalent question is: which of the other three models yields the "same" $\widehat{\text{OR}}$ as obtained for the GS model? A quick glance at the table indicates that the $\widehat{\text{OR}}$ estimate for the GS model is somewhat higher than the estimates for the other three models, *suggesting that only the GS model controls for confounding*.

Change of Estimate Results: 10% Rule

Model:	I (GS)	II	III	IV
*V*s in model	AGE,GEN	AGE	GEN	Neither
$\widehat{\text{OR}}$	26.2430	22.3606	23.5706	21.9881
Within 10% of GS?	-	No	No	No

Note: ±10% of 26.2430: (23.6187, 28.8673)

Moreover, if we use a "change-of-estimate" rule of 10%, we find that none of models II, III, or IV have an $\widehat{\text{OR}}$ within 10% of the $\widehat{\text{OR}}$ of 26.2430 for the GS model (I), although model III comes very close.

Only GS model controls confounding

Model at this point contains

$$E_1, E_2, \underbrace{V_1, \text{ and } V_2}_{\text{can't drop}}$$

haven't yet addressed

This result indicates that the only model that controls for confounding is the GS model. *That is, we cannot drop either AGE or GENDER from the model*.

Model A Output −2 1n L = 279.317

```
Analysis of maximum likelihood estimates
        Param      DF Estimate  Std Err  ChiSq    Pr > ChiSq
        Intercept  1  -5.0583   0.7643   43.8059  <.0001
Es  { PREVHOSP     1   1.4855   0.4032   13.5745  0.0002
    { PAMU         1   1.7819   0.3707   23.1113  <.0001
Vs  { AGE          1   0.0353   0.0092   14.7004  0.0001
    { GENDER       1   0.9329   0.3418    7.4513  0.0063
```

We therefore have decided that both *V*s need to stay in the model, but we have not yet addressed the *E*s in the model.

Wald for
E_1 (PREVHOSP): $P = 0.0002$
E_2 (PAMU): $P < 0.0001$

The only other variable that we might consider dropping at this point is E_1 or E_2, provided we decide that one of these is nonsignificant, controlling for the other. However, on inspection of the output for this model, shown again at the left, we find that the Wald statistic for E_1 is significant ($P = 0.0002$), as is the Wald statistic for E_2 ($P < 0.0001$).

EXAMPLE (continued)

Options A and B (continued)

Cannot drop PREVHOSP or PAMU

Thus, based on these Wald statistics, we cannot drop either variable from the model (and similar conclusions from LR tests).

Using Options A or B:

⇓

No-Interaction Model A is best model

$$\text{Logit } P(\mathbf{X}) = \alpha + \beta_1 E_1 + \beta_2 E_2 + \gamma_1 V_1 + \gamma_2 V_2$$

Consequently, using *Options* **A** *or* **B**, our *best model* is the (reduced) no-interaction model A, which we have called the Gold Standard model.

$$\mathbf{X}^* = (\underset{\text{yes}}{E_1{}^* = 1}, \underset{\text{yes}}{E_2{}^* = 1})$$

$$\text{vs. } \mathbf{X} = (\underset{\text{no}}{E_1 = 0}, \underset{\text{no}}{E_2 = 0})$$

$$\begin{aligned}\text{OR}_{\text{model A}} &= \exp[\beta_1(1-0) \\ &\quad + \beta_2(1-0)] \\ &= \exp[\beta_1 + \beta_2]\end{aligned}$$

For this model, then, the OR that compares a subject **X*** who is positive (i.e., yes) for both *E*s with a subject **X** who is negative (i.e., no) for both *E*s simplifies to the exponential formula shown at the left.

$$\begin{aligned}\widehat{\text{OR}} &= \exp[\hat{\beta}_1 + \hat{\beta}_2] \\ &= \exp[1.4855 + 1.7819] = \boxed{\mathbf{26.2415}}\end{aligned}$$

95% CI: **(11.5512, 59.6146)**

Below this formula, at the left, we show the estimated OR and a 95% confidence interval around this odds ratio.

Conclusion from **Options A** and **B**: **Very strong (but highly variable) combined effect of PREVHOSP and PAMU**

These results show that there is a very strong and significant (but highly variable) effect when comparing MRSA models with **X*** and **X**.

$$\left\{\begin{aligned}&\widehat{\text{OR}}_{E_1|E_2,V_1,V_2} = \exp[\hat{\beta}_1] = \exp[1.4855] \\ &\qquad = \boxed{\mathbf{4.417}} \\ &95\%\ \text{CI}_{E_1|E_2,V_1,V_2} = \mathbf{[2.2004,\ 9.734]}\end{aligned}\right.$$

$$\left\{\begin{aligned}&\widehat{\text{OR}}_{E_2|E_1,V_1,V_2} = \exp[\hat{\beta}_2] = \exp[1.7819] \\ &\qquad = \boxed{\mathbf{5.941}} \\ &95\%\ \text{CI}_{E_2|E_1,V_1,V_2} = \mathbf{[2.873,\ 12.285]}\end{aligned}\right.$$

Alternatively, we might wish to compute the odds ratios for the effects of each *E* variable, separately, controlling for the other *E* and the two *V* variables. The results are shown at the left and can also be obtained using the output for reduced model **A** shown earlier.

Options A or B: Additional conclusions

Both PREVHOSP and PAMU have moderately strong and significant individual effects.

From these results, we can conclude from using *Options* **A** or **B** that both PREVHOSP and PAMU have moderately strong and significant individual effects (ORs of 4.417 and 5.941, respectively) when controlling for the other three variables in the final model, i.e., no-interaction model **A**.

EXAMPLE (continued)

Option A: Overall (chunk) interaction, then, in order. *EW*s, *EE*s, *V*s, and *E*s

Option B: Assess *EW*s first, then, in order *EE*s, *V*s, and *E*s

Option C: Assess *EW*s first, then, in order, *V*s, *EE*s, and *E*s

Recall that both *Options* **A** *and* **B** assessed interaction of *EW*s and *EE*s before considering confounding and precision, where *Option* **A** used an overall (chunk) test for interaction and *Option* **B** did not. We are now ready to consider *Option* **C**, which assesses interactions involving *EW*s first, then confounding and precision (i.e., the *V*s), after which *EE*s and finally *E*s are evaluated.

Reduced Model B (w/o *EW* terms):

$$\text{Logit P}(\mathbf{X}) = \alpha + (\beta_1 E_1 + \beta_2 E_2) + (\gamma_1 V_1 + \gamma_2 V_2) + \delta^* E_1 E_2$$

Note: Reduced model **B** preferred to full interaction model

Since all three Options, including *Option* **C**, assess *EW*s before *EE*s, *V*s, and *E*s, we have already determined the results for the *EW*s. That is, we can drop all the *EW*s, which yields reduced model **B**, as shown again at the left.

Model B Output: –21n L = 277.667

```
Analysis of maximum likelihood estimates
         Param      DF  Estimate  Std Err   ChiSq   Pr > ChiSq
         Intercept  1   -4.9038   0.7649   41.1020  <.0001
Es  {PREVHOSP       1    1.0503   0.5257    3.9922  0.0457
    {PAMU           1    0.9772   0.7589    1.6583  0.1978
Vs  {AGE            1    0.0357   0.0092   15.0206  0.0001
    {GENDER         1    0.9778   0.3464    7.9660  0.0048
EE  {PRHPAM         1    1.0894   0.8733    1.5562  0.2122
```

The corresponding (edited) output for model **B** is shown again here. This model retains the *EE* product term PRHPAM (= E_1E_2), which using *Option* **C**, will not be considered for exclusion until we address confounding for AGE and GENDER (i.e., the *V*s).

Confounding:

Does $\widehat{\text{OR}}$ meaningfully change when AGE and/or GENDER are dropped?

To assess confounding, we need to determine whether the estimated OR meaningfully changes (e.g., by more than 10%) when either AGE or GENDER or both are dropped from the model. Here, the gold standard (GS) model is the reduced model **B**, which contains the E_1E_2 term.

GS model: reduced model **B** above

$$\text{OR}_{\text{GS}(\mathbf{B})} = \exp[\beta_1(E_1{}^* - E_1) + \beta_2(E_2{}^* - E_2) + \delta^*(E_1{}^*E_2{}^* - E_1E_2)],$$

where $\mathbf{X}^* = (E_1{}^*, E_2{}^*)$ and $\mathbf{X} = (E_1, E_2)$ are two specifications of the two *E*s

The formula for the odds ratio for the GS model is shown at the left, where (E_1*, E_2*) and (E_1, E_2) denote two specifications of the two exposures PREVHOSP (i.e., E_1) and PAMU (i.e., E_2). This formula contains three parameters: β_1, β_2, and δ^*.

EXAMPLE (continued)

Option C (continued)

Must specify **X*** and **X:**

$$\mathbf{X}^* = (E_1{}^* = \underset{\text{yes}}{1},\ E_2{}^* = \underset{\text{yes}}{1})$$

$$\text{vs. } \mathbf{X} = (E_1 = \underset{\text{no}}{0},\ E_2 = \underset{\text{no}}{0})$$

As previously noted (for *Option* **A**), there are several ways to specify **X*** and **X**. Here, again, we will choose to compare a subject **X*** who is positive (i.e., yes) for both *E*s with a subject **X** who is negative (i.e., no) for both *E*s.

$$\begin{aligned}\text{OR}_{\text{GS}(\mathbf{B})} &= \exp[\beta_1(1-0) + \beta_2(1-0) \\ &\quad + \hat{\delta}^*([1\times 1] - [0\times 0])] \\ &= \exp[\beta_1 + \beta_2 + \delta^*]\end{aligned}$$

Based on the above choices, the OR formula for our **GS** reduced model **B** simplifies, as shown here.

Table of $\widehat{\text{OR}}$s (check confounding)

*V*s in model	AGE,GEN	AGE	GEN	Neither
$\widehat{\text{OR}}$	$\widehat{\text{OR}}_{\text{I}^*}$	$\widehat{\text{OR}}_{\text{II}^*}$	$\widehat{\text{OR}}_{\text{III}^*}$	$\widehat{\text{OR}}_{\text{IV}^*}$

To assess confounding, we must once again (as with *Option* **A**) consider a table of $\widehat{\text{OR}}$s, as shown at the left.

Model choices:

$$\text{I}^*.\ \text{Logit P}_{\text{I}^*}(\mathbf{X}) = \alpha + \beta_1E_1 + \beta_2E_2 + \gamma_1V_1 + \gamma_2V_2 + \delta^*E_1E_2$$

$$\text{II}^*.\ \text{Logit P}_{\text{II}^*}(\mathbf{X}) = \alpha + \beta_1E_1 + \beta_2E_2 + \gamma_1V_1 + \delta^*E_1E_2$$

$$\text{III}^*.\ \text{Logit P}_{\text{III}^*}(\mathbf{X}) = \alpha + \beta_1E_1 + \beta_2E_2 + \gamma_2V_2 + \delta^*E_1E_2$$

$$\text{IV}^*.\ \text{Logit P}_{\text{IV}^*}(\mathbf{X}) = \alpha + \beta_1E_1 + \beta_2E_2 + \delta^*E_1E_2$$

To complete the above table, we need to fit the four models shown at the left. The first model (I*), which we have already described, is the Gold Standard (**GS**(**B**)) model containing PREVHOSP, PAMU, AGE, GENDER, and PRHPAM. The other three models exclude GENDER, AGE, or both from the model.

OR formula ($E_1{}^* = 1, E_2{}^* = 1$) vs. ($E_1 = 0, E_2 = 0$) for all four models:

$$\text{OR} = \exp[\beta_1 + \beta_2 + \delta^*]$$

Since all four models involve the same two *E* variables, the general formula for the OR that compares a subject who is exposed on both *E*s ($E_1{}^* = 1,\ E_2{}^* = 1$) vs. a subject who is not exposed on both *E*s ($E_1 = 0,\ E_2 = 0$) has the same algebraic form for each model, including the **GS**(**B**) model.

However, $\hat{\beta}_1, \hat{\beta}_2$, and $\hat{\delta}^*$ likely differ for each model

However, since, the models do not all have the same predictors, the estimates of the regression coefficients are likely to differ somewhat.

EXAMPLE (continued)

Option C (continued)

Estimated Regression Coefficients and ORs

Model:	I*(GS)	II*	III*	IV*
*V*s in model	AGE, GEN	AGE	GEN	Neither
$\hat{\beta}_1$	1.0503	1.1224	1.2851	1.2981
$\hat{\beta}_2$	0.9772	1.0021	0.8002	0.8251
$\hat{\delta}^*$	1.0894	0.8557	0.9374	0.8398
$\widehat{OR}$	22.5762	19.6918	20.5467	19.3560

At the left, we show for each model, the values of these three estimated regression coefficients together with their corresponding OR estimates.

Confounding (**Option C**):
Which models have "same" $\widehat{OR}$ as **GS** model?

From this information, we must decide whether any one or more of models II*, III*, and IV* yields the "same" $\widehat{OR}$ as obtained for the **GS** model (I*).

$\widehat{OR}$ for **GS** is highest
⇓
Only **GS** model controls confounding

Notice, first, that the OR estimate for the **GS** model (22.5762) is somewhat higher than the estimates for the other three models, suggesting that only the **GS** model controls for confounding.

Change of Estimate Results: 10% Rule

Model:	I* (GS)	II*	III*	IV*
*V*s in model	AGE, GEN	AGE	GEN	Neither
$\widehat{OR}$	**22.5762**	19.6918	**20.5467**	19.3560
Within 10% of GS?	–	No	**Yes**	No

Note: ±10% of 22.5762: (20.3186, 24.8338)

However, using a "change-of-estimate" rule of 10%, we find that the $\widehat{OR}$ (20.5467) for model III*, which drops AGE but retains GENDER, is within 10% of the $\widehat{OR}$ (22.5762) for the **GS** model. This result suggests that there are two candidate models (I* and III*) that control for confounding.

Two alternative conclusions:

(a) Only **GS** model controls confounding

(b) **GS** model (I*) and model III* both control confounding

From the above results, we must decide at this point which of two conclusions to draw about confounding: (a) the only model that controls for confounding is the **GS** model; or (b) both the **GS** model (I*) and model III* control for confounding.

EXAMPLE (continued)

Option C (continued)

Suppose decide only **GS(B)** control confounding

⇓

Model at this point contains

$$E_1, E_2, \underbrace{V_1, V_2}_{\text{can't drop}} \text{ and } E_1E_2,$$

have not yet addressed

Suppose we decide that only the **GS** model controls for confounding. Then, we cannot drop either AGE or GENDER from the model. We therefore have decided that both *V*s need to stay in the model, but we have not yet addressed the *E*s in the model.

Next step: test E_1E_2:

Wald χ^2 (reduced model **B**) $= 1.5562, P = 0.2122$ (n.s.)

$$\text{LR} = -2\ln \text{L}_{\text{model } \mathbf{A}} - 2\ln \text{L}_{\text{model } \mathbf{B}})$$
$$= 279.317 - 277.667 = 1.650 \sim \chi^2_{1\text{ df}}$$
$$(P = 0.1989)$$

For the next step, we would test whether the E_1E_2 product term is significant.

From our output for reduced model **B** given previously, we find that the Wald test for the PRHPAM term (i.e., E_1E_2) is not significant ($P = 0.2122$). The corresponding LR test is obtained by comparing −2lnL statistics for *reduced models* **A** and **B**, yielding a LR statistic of 1.650, also nonsignificant.

No-interaction Model **A**:

$$\text{Logit P}(\mathbf{X}) = \alpha + (\beta_1 E_1 + \beta_2 E_2) + (\gamma_1 V_1 + \gamma_2 V_2),$$

where $V_1 = C_1 =$ AGE,
$V_2 = C_2 =$ GENDER
$E_1 =$ PREVHOSP,
$E_2 =$ PAMU

We can now reduce our model further by dropping the E_1E_2 term, which yields the no-interaction model **A**, shown at the left.

Recall : **Options A and B** ⇒ Model **A** is best

Option C : only **GS** model controls confounding

⇓

Model **A** is best

Recall that Model **A** was chosen as the best model using *Options* **A** and **B**. Consequently, using *Option* **C**, if we decide that the only model that controls confounding is the **GS**(**B**) model (I* above), then our best model for *Option* **C** is also Model **A**.

Alternative decision about confounding for **Option C**

⇓

2 candidate models control confounding:
Model I*: **GS(B)**
Model III*: (AGE dropped)

The above conclusion (i.e., Model **A** is best), nevertheless, resulted from the decision that only the **GS(B)** model controlled for confounding. However, we alternatively allowed for two candidate models, the **GS(B)** model (I*) and model III*, which dropped AGE from the model, to control for confounding.

How to decide between models?
Answer: **Precision**

If we decide to consider model III* in addition to the **GS(B)** model, how do we decide between these two models? The answer, according to the modeling strategy described in Chap. 7, is to consider *precision*.

EXAMPLE (continued)

Option C (continued)

OR formula for Models I* and III*:

$$\text{OR} = \exp[\beta_1(E_1{}^* - E_1) + \beta_2(E_2{}^* - E_2) + \delta^*(E_1{}^*E_2{}^* - E_1E_2)],$$

where $\mathbf{X^*} = (E_1{}^*, E_2{}^*)$ and $\mathbf{X} = (E_1, E_2)$ are two specifications of the two Es

Since both Models I* and III* include the interaction term E_1E_2, the OR formula has the same structure for both models (shown again at the left).

Precision ⇒ computing CIs for the OR for Models I* and III*

To evaluate precision for each odds ratio, we must therefore compute (say, 95%) confidence intervals (CIs) for the OR for each model.

CI depends on how we specify **X*** and **X**:
Our focus again:

$$\mathbf{X^*} = (1, 1) \text{ vs. } \mathbf{X} = (0, 0)$$
$$\Downarrow$$
$$\text{OR} = \exp[\beta_1 + \beta_2 + \delta^*]$$

The CI limits for each OR will depend on how we specify **X*** and **X**. As we did for confounding, we again focus on comparing a subject **X*** who is positive (i.e., yes) for both Es with a subject **X** who is negative (i.e., no) for both Es. The OR formula simplifies as shown at the left.

Table of ORs and CIs for Models I* and III*

	$\widehat{\text{OR}}$	95% CI for OR
Model I* (GS(**B**))	22.5762	**(10.0175,50.8871)**
Model III* (w/o AGE)	20.5467	**(9.4174,44.8250)**

To assess precision, therefore, we must now consider a table that gives the (95%) CI for the OR for each model, and then decide whether or not precision is gained when AGE is dropped from the GS model. The resulting table is shown at the left.

CI width

Model I*	50.8871 − 10.0175 = **40.8696**
Model III*	44.8250 − 9.4174 = **35.4076**

From the above results, we can see that, although both models give wide (i.e., imprecise) confidence intervals, Model III* has a tighter confidence interval than Model I*.

Better model: Model III*

$$\text{Logit } P_{\text{III}^*}(\mathbf{X}) = \alpha + (\beta_1E_1 + \beta_2E_2) + \gamma_1V_2 + \delta^*E_1E_2$$

(same OR but better precision than **GS**)

Therefore, we suggest that Model III* be chosen as the "better" model, since it gives the "same" (within 10%) OR estimate and provides more precision.

Model III* at this point contains

$$E_1, E_2, V_2, \text{ and } E_1E_2,$$

haven't yet addressed

At this point, using model III*, we have decided to drop V_1 = AGE from our initial model. Nevertheless, we have not yet addressed the Es in the model.

EXAMPLE (continued)

Option C (continued)

Model III* Output

```
Analysis of maximum likelihood estimates
                           Std
Param      DF  Estimate    Err     ChiSq    Pr > ChiSq
Intercept  1   −2.6264   0.4209  38.9414  <.0001
PREVHOSP   1    1.2851   0.5107   6.3313  0.0119
PAMU       1    0.8002   0.7317   1.1960  0.2741
GENDER     1    0.4633   0.3066   2.2835  0.1308
PRHPAM     1    0.9374   0.8432   1.2358  0.2663 n.s
```

Model C :
Logit $P(\mathbf{X}) = \alpha + (\beta_1 E_1 + \beta_2 E_2) + \gamma_2 V_2$

Can we drop E_1 or E_2 from Model C?

Model C Output

```
Analysis of maximum likelihood estimates
Param     DF Estimate  Std Err  ChiSq    Pr > ChiSq
Intercept 1  -2.7924   0.4123  45.8793  <.0001
PREVHOSP  1   1.6627   0.3908  18.1010  <.0001
PAMU      1   1.4973   0.3462  18.7090  <.0001
GENDER    1   0.4335   0.3030   2.4066  0.1525
```

Cannot drop either E_1 or E_2

Option C conclusion:
Model C is best model

$\mathbf{X}^* = (E_1{}^* = 1, E_2{}^* = 1)$ vs.
$\mathbf{X} = (E_1 = 0, E_2 = 0)$
$\text{OR}_{\text{Model C}} = \exp[\beta_1 + \beta_2]$

$$\widehat{\text{OR}}_{\text{Model C}} = \exp[\hat{\beta}_1 + \hat{\beta}_2] = \exp[1.6627 + 1.4973] = 23.5708$$

95% CI: **(10.7737, 51.5684)**

For the next step, we would test whether the E_1E_2 product term is significant. Using the output for Model III* (shown at the left), we find that the Wald test for the PRHPAM term (i.e., E_1E_2) is not significant ($P = 0.2663$). The corresponding LR test is also not significant.

We can now reduce our model further by dropping the E_1E_2 term, which yields the reduced Model C, shown at the left.

The only other variables that we might consider dropping at this point are E_1 or E_2, provided one of these is not significant, controlling for the other.

However, on inspection of the output for this model, shown at the left, we find that the Wald statistic for E_1 is highly significant ($P < 0.0001$), as is the Wald statistic for E_2 ($P < 0.0001$). Thus, based on these Wald statistics, we cannot drop either E variable from the model (and similar conclusions from LR tests).

Consequently, if we decide to use *Option* C, and we allow Models I* and Models III* to be candidate models that control for confounding, then our *best model* is given by Model **C**. To make this choice, we considered precision as well as significance of the E in the model.

For this model, then, the OR that compares a subject **X*** who is positive (i.e., yes) for both *E*s with a subject **X** who is negative (i.e., no) for both *E*s simplifies to the exp formula shown at the left.

Below this formula, we show the estimated OR and a 95% confidence interval around this odds ratio, which indicates a very strong and significant (but highly variable) effect.

EXAMPLE (continued)

Best Model Summary: Options A, B, C

Options A and B (same result):
Model **A**: contains PREVHOSP, PAMU, AGE, and GENDER

Option C:
Model **C**: contains PREVHOSP, PAMU, and GENDER

Summarizing the results we have obtained from the above analyses on the MRSA data, we have found *two* different final choices for the best model shown at the left depending on three approaches to our modeling strategy, *Options* **A** and **B** (same result) and *Option* **C**.

Model A Output (Best: Options A and B)

```
Analysis of maximum likelihood estimates
Param      DF  Estimate  Std Err  ChiSq    Pr > ChiSq
Intercept  1   −5.0583   0.7643   43.8059  <.0001
PREVHOSP   1    1.4855   0.4032   13.5745  0.0002
PAMU       1    1.7819   0.3707   23.1113  <.0001
AGE        1   0.0353    0.0092   14.7004  0.0001
GENDER     1   0.9329    0.3418    7.4513  0.0063
```

The outputs for the two "best" models are shown here.

Both models are no-interaction models, and they both contain the main effects of two highly significant *E* variables, PREVHOSP and PAMU.

Model C Output (Best: Option C)

```
Analysis of maximum likelihood estimates
Param      DF  Estimate  Std Err  ChiSq    Pr > ChiSq
Intercept  1   −2.7924   0.4123   45.8793  <.0001
PREVHOSP   1    1.6627   0.3908   18.1010  <.0001
PAMU       1    1.4973   0.3462   18.7090  <.0001
GENDER     1    0.4335   0.3030   2.4066   0.1525
```

The estimated coefficients of PREVHOSP and PAMU differ somewhat for each model. The estimate for PREVHOSP is 1.4855 for Model **A** whereas it is 1.6627 for Model **C**. The estimate for PAMU is 1.7819 for Model **A** compared with 1.4973 for Model **C**.

		$\widehat{\text{ORs}}$	
	PREVHOSP	**PAMU**	**COMBINED**
Model	$\exp[\hat{\beta}_1]$	$\exp[\hat{\beta}_2]$	$\exp[\hat{\beta}_1+\hat{\beta}_2]$
A	**4.417**	**5.941**	**26.242**
C	**5.274**	**4.470**	**23.571**

OR estimates from each model are shown in the table at the left. Both models show moderately strong effects for each *E* variable and a very strong effect when comparing $\mathbf{X}^* = (E_1 = 1, E_2 = 1)$ with $\mathbf{X} = (E_1 = 0, E_2 = 0)$. However, the effect of PREVHOSP is 16*% lower* in Model A than in Model **C**, whereas the effect of PAMU 25*% higher* in Model **A** than in Model **C**.

MRSA example: **Options A, B, and C**
⇓
Similar, slightly different numerical conclusions

We see, therefore, that for our MRSA example, modeling strategy *Options* **A, B**, and **C** give similar, but slightly different conclusions involving two *E* variables.

In general: No guarantee for same conclusions

In general, as shown by this example, there is no guarantee that these three options will always yield the same conclusions. Therefore, the researcher may have to decide which option he/she prefers and/or which conclusion makes the most (biologic) sense.

General form of Initial Model

$$\text{Logit P}(\mathbf{X}) = \alpha + \sum_{i=1}^{q}\beta_i E_i + \sum_{j=1}^{p_1}\gamma_j V_j + \sum_{i=1}^{q}\sum_{k=1}^{p_2}\delta_{ik}E_i W_k + \sum_{i=1}^{q}\sum_{\substack{i'=1\\ i\neq i'}}^{q}\delta^*_{ii'}E_i E_{i'}$$

In summary, we recommend that the initial model has the general form shown at the left. This model involves *E*s, *V*s, *EW*s, and *EE*s, so there are two types of interaction terms to consider.

Modeling Strategy Summary: Several *E*s

Step 1: Define initial model (above formula)

Step 2: Assess interaction
Option A: Overall chunk test + Options B or C
Option B: Test *EW*s, then *EE*s
Option C: Test *EW*s, but assess *V*s before *EE*s

Step 3: Assess confounding and precision (*V*s)
Option A and B (cont'd):
*V*s after *EW*s and *EE*s
Option (cont'd):
*V*s after *EW*s, but prior to *EE*s

Step 4: Test for nonsignificant *E*s if not components of significant *EE*s

We then recommend assessing interaction, first by deciding whether to do an overall chunk test, then testing for the *EW*s, after which a choice has to be made as to whether to test for the *EE* terms prior to or subsequent to assessing confounding and precision (involving the *V*s).

The resulting model can then be further assessed to see whether any of the *E* terms in the model can be dropped as nonsignificant.

Special Cases: Several *E*s

There are two special cases that we now address:

(a) **All *V*s are controlled as main effects,**
i.e., confounding and precision for *V*s not considered

What if you decide to control for all *V*s as main effects (without assessing confounding and/or precision)?

Modeling Strategy: All *V*s controlled

Step 1: Define initial model (above formula)

Step 2: Assess Interaction
Option A: Overall chunk test + Options B
Option B: Test *EW*s, then *EE*s

Step 4: Test for nonsignif *E*s if not components of significant *EE*s

In this case (**a**), we only need to consider Options A and B, so that Step 3 of our previously described strategy can be omitted.

EXAMPLE

MRSA Initial Model, Special case(a)

$$\begin{aligned}\text{Logit P}(\mathbf{X}) = \alpha &+ (\beta_1 E_1 + \beta_2 E_2)\\ &+ (\gamma_1 V_1 + \gamma_2 V_2)\\ &+ (\delta_{11} E_1 W_1 + \delta_{12} E_1 W_2\\ &+ \delta_{21} E_2 W_1 + \delta_{22} E_2 W_2)\\ &+ \delta^* E_1 E_2\end{aligned}$$

Model **A**: Final Model

$$\begin{aligned}\text{Logit P}(\mathbf{X}) = \alpha + (\beta_1 E_1 + \beta_2 E_2) + (\gamma_1 V_1\\ + \gamma_2 V_2)\end{aligned}$$

For example, using the MRSA data, the initial model, shown again at the left contains two *E*s, two *V*s, 4 *EW*s and one *EE* term.

When previously applying Options A and B to this model, we dropped all interaction terms, resulting in reduced model A shown at the left. If we decide in advance to control for both *V*s, then this is our final model, since both *E*s were significant in this model.

(b) **The model contains only *E*s and *EE*s, but no *C*s**
(i.e., no *V*s or *W*s)
"Hypothesis Generating" Model

As a second special case, *what if our model contains only* E*s, so there are no Cs to control?* This case is often referred to as a "hypothesis generating" model, since we are essentially assuming that we have limited knowledge on all possible predictors and that no risk factors have been established.

General Model: Only *E*s and *EE*s

$$\text{Logit P}(\mathbf{X}) = \alpha + \sum_{i=1}^{q} \beta_i E_i + \sum_{i=1}^{q} \sum_{\substack{i'=1 \\ i \neq i'}}^{q} \delta^*_{ii'} E_i E_{i'}$$

In this case (**b**), our general model takes the simplified form shown at the left.

Modeling Strategy: All *E*s, no *C*s

Step 1: Define initial model (above formula)

Step 2: Assess interaction involving *E*s
Option A*: Overall chunk test for *EE*s, followed by backward elimination of *EE*s

Step 4: Test for nonsignif *E*s if not components of significant *EE*s

For this model, we recommend a correspondingly simplified strategy as shown at the left that involves statistical testing only, first for *EE* terms, and then for *E*s that are not components of significant *EE*s. In terms of the options we previously described, we only need to consider a modified version of Option A, and that Step 3, once again, can be omitted.

EXAMPLE

MRSA example Initial Model, Special case (b)

$\text{Logit P}(\mathbf{X}) = \alpha + \beta_1 E_1 + \beta_2 E_2 + \delta^* E_1 E_2$

Applying this situation to our MRSA data, the initial model (w/o the *C*s) is shown at the left.

Final model: All *E*s, no *C*s:

$\text{Logit P}(\mathbf{X}) = \alpha + \beta_1 E_1 + \beta_2 E_2$,
where E_1 = PREVHOSP and
E_2 = PAMU

Testing H_0: $\delta^* = 0$ in this model yields a nonsignificant result (data not shown), and the final model (since individual *E*s cannot be dropped) is the no-interaction model shown here.

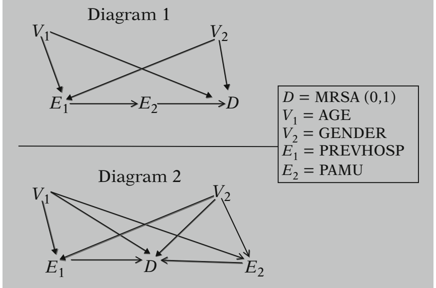
One other issue: specifying the initial model

(MRSA example)

We now address one other issue, which concerns how to specify the initial model. We describe this issue in the context of the MRSA example.

EXAMPLE

Possible Causal Diagrams for MRSA Study



At the left, we consider two possible *causal diagrams* for the MRSA data.

Diagram 1 ⇒ PAMU intervening variable;
AGE and GENDER confounders

Diagram 1 indicates that PAMU (i.e., E_2) is an intervening variable in the causal pathway between PREVHOSP (i.e., E_1) and MRSA outcome, and that AGE and GENDER (i.e., V_1 and V_2) are confounders of the relationship between PREVHOSP and MRSA.

EXAMPLE (continued)

Diagram 2 ⇒ PREVHOSP and PAMU independent risk factors; AGE and GENDER confounders

Diagram 2 indicates that PREVHOSP and PAMU are independent risk factors for MRSA outcome, and that AGE and GENDER are confounders of both PREVHOSP and PAMU.

Diagram 2 appropriate ⇒ initial model containing both E_1 and E_2 is justified

The initial model that we considered in our analysis of the MRSA data, containing both PREVHOSP and PAMU in the model as *E* variables, can be justified if we decide that Diagram 2 is a correct representation of the causal pathways involved.

Diagram 1 appropriate ⇒ initial model should not contain both PREVHOSP and PAMU

In contrast, if we decide that Diagram 1 is more appropriate than Diagram 2, we should not put both PREVHOSP and PAMU in the same model to assess their joint or separate effects.

i.e., PAMU intervening variable

$$\Downarrow$$

$$\text{Logit P}(\mathbf{X}) = \alpha + \beta_1 E_1 + \gamma_1 V_1 + \gamma_2 V_2 + \delta_{11} E_1 V_1 + \delta_{12} E_1 V_2$$

In other words, if PAMU is an intervening variable, we should consider a model involving only one *E* variable, preferably PREVHOSP. An example of such a model, which controls for AGE and GENDER and allows for interaction effects, is shown at the left.

The moral: Causal diagram can influence choice of initial model

Thus, as mentioned previously in Chap. 6, the choice of the initial model can be influenced by the causal diagram considered most appropriate for one's data.

III. Screening Variables

Scenario:

Logistic Model
$E(0,1)$ vs. $D(0,1)$
$C_1, C_2, \ldots, C_p$
"large" p

Desired initial model:

$$\text{Logit P}(\mathbf{X}) = \alpha + \beta E + \sum_{j=1}^{p} \gamma_j C_j + \sum_{j=1}^{p} \delta_j E C_j$$

Follow hierarchical BW elimination strategy (Chap. 7)

However, suppose:

- Computer program does not run
  or
- Fitted model unreliable ("large" p)

What do you do?

In this section, we address the following scenario: Suppose you wish to fit a binary logistic model involving a binary exposure and outcome variables E and D controlling for the potential confounding and effect-modifying effects of a "large" number of variables C_j, j = 1, 2, ... , p that you have identified from the literature.

You would like to begin with a model containing E, the main effects of each C_j, and all product terms of the form $E \times C_j$, and then follow the hierarchical backward elimination strategy described in Chap. 7 to obtain a "best" model.

However, when you run a computer program (e.g., SAS's Proc Logistic) to fit this model, you find that the model does not run or you decide that, even if the model runs, the resulting fitted model is too unreliable because of the large number of variables being considered. What do you do in this situation?

OPTIONS (large-number-of-variables problem)

1. **Screening:**
   - Exclude some C_j one-at-a-time
   - Begin again with reduced model
2. **Collinearity diagnostics on initial model:**
   - Exclude some C_j and/or $E \times C_j$ strongly related to other variables in the model
3. **Forward algorithm for interactions:**
   - Start with E and all C_j, $j = 1, \ldots, p$
   - Sequentially add significant $E \times C_j$

There are several possible options:

1. Use some kind of "screening" technique to exclude some of the C_j variables from the model one-at-a-time, and then begin again with a reduced-sized model that you hope is reasonably reliable and/or at least will run.

2. Use "collinearity" diagnostic methods starting with the initial model to exclude variables (typically product terms) that are strongly related to other variables in the model.

3. Use a forward regression algorithm that starts with a model containing all main effect C_j terms and proceed to sequentially add statistically significant product terms.

4. **Backward for Cs, then forward for $E \times C_j$:**
   - Start with E and all C_j, $j = 1, \ldots, p$
   - Sequentially drop nonsignif. C_j
   - Sequentially add $E \times C_j$ for remaining C_j

4. Start with a model containing all C_j terms, proceed backward to eliminate nonsignificant C_j terms, and then sequentially add statistically significant product terms among the remaining C_j terms.

COMMENTS/CRITICISMS OF OPTIONS:

Option 2: next section.

Option 2 above will be described in the next section. It cannot be used, however, if initial model does not run.

Option 3:

+ Starts with small-sized model

− Can still be unreliable if large number of Cs

− Interactions not assessed simultaneously

Option 3 has the advantage of starting with a small-sized model, but has two disadvantages: the model may still have reliability problems if there are a "large" number of Cs, and the forward approach to assess interaction does not allow all interaction terms to be assessed simultaneously as with a backward approach.

Option 4:

+ Frequently used in practice (??)

− Inappropriately uses statistical testing to exclude potential confounders

− Questionably excludes Cs before assessing $E \times C$s

Option 4, which is frequently used in practice, can be *strongly criticized* because it uses statistical testing to determine whether potential confounders C_j should stay in the model, whereas statistical testing should not be used to assess confounding. Furthermore, option 4 excludes potential confounders prior to assessing interaction, whereas interaction should be assessed before confounding.

Screening:
Good ways and questionable ways

We now return to describe Option 1: screening. As we will describe below, there are good ways and questionable ways to carry out screening.

Purpose:

- Reduce number of predictors
- Obtain a reliable and interpretable final model

The purpose of screening is to reduce the number of predictors being considered so that a reliable and interpretable final model can be obtained to help answer study questions of interest.

The two primary drawbacks of screening are:

Drawbacks of screening:

1. No simultaneous assessment of all *E*s and *C*s.
2. No guarantee final model contains all "relevant" variables

1. Does not accomplish simultaneous assessment of all exposure and control variables recommended from the literature or conceptualization of one's research question.
2. No guarantee that one's final model contains all the relevant variables of interest, although there is no such guarantee for any modeling strategy.

General screening situation:

n subjects, k predictors (X_i, $i = 1, \ldots, k$)

k "large enough" to require screening

Consider the following *general screening situation*: Your dataset contains n subjects and k predictors, and you decide k is large enough to warrant some kind of screening procedure to reduce the number of predictors in your initial model.

Method 0:

- **Consider predictors one-at-a-time**
- **Screen-out those X_i not significantly associated with *D***

A typical approach (let's call it *Method 0*) is to screen-out (i.e., remove from one's initial model) those variables that are not individually significantly associated with the (binary) outcome.

Questions about Method 0:

Q1. Any criticism?

Q1. Is there anything that can be criticized about Method 0?

Q2. Depends on types of *X*s?
- Use if several *E*s and *C*s?
- Use if one *E* and several *C*s?
- Use if only *E*s?

Q2. Should the use of Method 0 depend on types of predictors? E.g., whether your predictors are a mixture of *E*s and *C*s, involve one *E* and several *C*s, or only involve *E*s?

Q3. How large k compared to n?
- $k = 10, n = 50$: 20%?
- $k = 10, n = 100$: 10%?
- $k = 10, n = 200$: 5%?

Q3. How large does k have to be *relative to n* in order to justify screening?

Q4. Other ways than Method 0?

Q4. Are there other ways (i.e., Methods A, B, C, ...) to carry out (one-at-a-time) screening and when, if at all, should they be preferred to the typical approach?

Q5. Collinearity and/or screening?

Q5. Where does collinearity assessment fit in with this problem?

Answers:

Q1. **Yes:**

- Statistical testing only (questionable)
- Does not consider confounding or interaction

Q1. Is there anything that can be criticized about Method 0?

Yes, Method 0 involves statistical testing only; it does not consider confounding or effect modification (interaction) when assessing variables one-at-a-time.

Assess confounding with *D*(0,1), *E*(0,1), one *C*:

Logit $\mathrm{P}(\mathbf{X}) = \alpha + \beta E$,
where $\mathrm{P}(\mathbf{X}) = \Pr(D=1|E)$

Logit $\mathrm{P}^*(\mathbf{X}) = \alpha^* + \beta^* E + \gamma^* C$,
where $\mathrm{P}^*(\mathbf{X}) = \Pr(D=1|E,C)$

To assess confounding involving binary disease *D*, binary exposure *E*, and a single potential confounder *C*, you need to fit two regression models (shown at left), one of which contains *E* and *C*, and the other of which contains only *E*.

Confounding:

$$\widehat{\mathrm{OR}}_{DE} = e^{\hat{\beta}} \overset{\boxed{\text{meaningfully different}}}{\neq} \widehat{\mathrm{OR}}_{DE|C} = e^{\hat{\beta}^*}$$

Confounding is present if we conclude that corresponding odds ratio estimates are meaningfully different for the two models.

Assess interaction with *D*(0,1), *E*(0,1), one *C*:

Logit $\mathrm{P}(\mathbf{X}) = \alpha + \beta E + \gamma C + \delta E \times C$,
where $\mathrm{P}(\mathbf{X}) = \Pr(D=1|E,C,E \times C)$

To assess interaction involving binary disease *D*, binary exposure *E*, and a single potential confounder *C*, we need to fit the following logistic regression model shown at the left that contains the main effects of *E* and *C* and the product term *E* × *C*.

H_0: $\delta = 0$

Wald $= \left(\hat{\delta}/s_{\hat{\delta}}\right)^2 \sim \chi^2_{1\text{ df}}$ under H_0

LR $= -2\ln L_R - (-2\ln L_F) \sim \chi^2_{1\text{ df}}$ under H_0

Interaction is then assessed by testing the null hypothesis that the coefficient (δ) of the product term is zero using either a Wald test or a likelihood ratio test (preferred), where the test statistic is chi square with 1 df under H_0.

Answers about Screening (continued)

Q2. **Yes: It depends.**

- Use Method 0 if several *E*s and *C*s? No
- Use Method 0 if one *E* and several *C*s? No
- Use Method 0 if only *E*s? Yes

Questionable when model considers *C*s

Q2. Should the use of Method 0 depend on types of predictors?
Yes, Method 0 makes most sense when the model only involves *E*s, i.e., no potential confounders (*C*s) and no corresponding changes in ORs are being considered. However, Method 0 is questionable whenever there are variables being controlled (*C*s).

Q3. **No clear-cut rules for "large *k*."**
However, need screening if initial model does not run.

Q3. How large does *k* have to be relative to *n* in order to justify screening?

There are no clear-cut rules, but you will become aware that screening should be considered if your initial model does not run (see next section on collinearity).

Q4. Other options for (1-at-a-time) screening? **Yes**

- Variations to assess confounding or interaction, e.g., use stratified analysis instead of logistic regression
- Such options needed if considering *C*s

Q4. Are there other ways to carry out (one-at-a-time) screening and when, if at all, should they be preferred to the typical approach?

There are several reasonable options for screening, all of which are variations of ways to assess possible confounding and/or effect modification involving covariates. Such options should be preferred whenever there is a mixture of *E*s and *C*s to be considered.

5. Collinearity?
√ Prior to screening and/or after screening

Q5. Where does collincarity assessment fit in with this problem?

Collinearity may be considered prior to screening or after screening is performed.

Initial model does not run
⇓
Cannot obtain collinearity diagnostics
⇓
Start with screening

If your initial model does not run, typical collinearity diagnostics (e.g., condition indices, to be described in the next section) cannot be obtained, so screening must be considered from the beginning.

Screening completed
⇓
Model may still be unreliable
⇓
Consider collinearity diagnostics

Also, once screening has been performed, collinearity assessment may determine that your reduced model (after screening) is still unreliable.

Question 2 (continued)
What if only *E*s (no *C*s considered)?
Answer: **Use Method 0.**

Following up on a previous question (2), suppose your starting model involves several *E*s but no *V*s or *W*s, i.e., no *C*s are considered. How do you carry out (one-at-a-time) screening for this situation?

Why? Confounding not an issue

We recommend using Method 0 here, because when there are no *C*s to consider, confounding is not an issue. Consequently, using statistical testing for one-at-a-time screening of *E*s is appropriate.

Initial model with *E*s and *C*s?
Screening procedure: **it depends!**
Option 1: Screen *C*s only without using Method 0
Option 2: Screen *C*s without using Method 0, and screen *E*s using Method 0

Suppose your initial model involves several *E*s and *C*s. How do you carry out screening for this situation? *The answer is, as is often the case, it depends!*
Option 1: You may decide to screen only *C*s, and then consider the *E*s during your modeling strategy process.
Option 2: If you have large numbers of *E* and *C* variables, you might screen both types of variables, making sure not to use Method 0 for the *C*s and using Method 0 for the *E*s.

EXAMPLE

Examples of Screening: Single E, C_1, C_2, . . . , C_{10}.

Four Scenarios: Which of these is "legitimate"?

i. Crude analyses relating D to each C_i identify only C_1 and C_4 to be significant predictors of D. Starting model then contains E, C_1, C_4, EC_1, and EC_4. Best model determined using hierarchical backward elimination approach (HBWE) outlined in Chap. 6

ii. Stratified analyses relating D to E and each C_i identify C_1 and C_4 to be individual confounders, and C_5 to be an effect modifier of the E, D effect. Starting model then contains E, C_1, C_4, C_5, EC_1, EC_4, and EC_5. Best model determined using HBWE.

iii. Crude analyses relating D to each C_i identify only C_1 and C_4 to be significant predictors of D. Starting model then contains E, C_1, and C_4. Backward elimination on Cs eliminates C_4, but retains C_1 (and E). Add interaction term EC_1. Best model determined using HBWE.

iv. Logistic regression models relating D to E and each C_i identify C_1 and C_4 to be individual confounders, and C_5 to be an effect modifier of the E, D effect. Starting model then contains E, C_1, C_4, C_5, EC_1, EC_4, and EC_5. Best model determined using HBWE.

We now provide a few simple examples to illustrate screening. We will begin by assuming that we have a single E variable and 10 C variables, C_1, C_2, . . . , C_{10}. At the left, we describe four different screening scenarios for this situation. Can you determine which of these scenarios corresponds to carrying out Method 0, and which represents what we have described above as a "legitimate" method of screening?

EXAMPLE (continued)

Answer:

Scenarios ii and iv "legitimate" screening

Scenarios i and iii incorrectly use significance tests to screen individual C_i.

Scenario iii questionably uses BW elimination on Cs before assessing interaction

The answer to the above question is that scenarios ii and iv represent "legitimate" methods of screening because both scenarios do not involve using a significance test of a crude effect between C_i and D. Scenario iii differs from i in that backward elimination is (questionably) performed on the Cs before interaction is assessed.

Summary about Method 0:

1. Does not assess confounding or interaction for individual C_i.
2. Makes most sense if model only involves Es.

Summarizing our main points about Method 0:

1. Method 0 does not consider confounding and/or interaction for predictors treated one-at-a-time.
2. Method 0 makes most sense when the model only involves Es, but is questionable with both Es and Cs being considered.

IV. Collinearity

Can some Xs be predicted by other Xs?

Collinearity concerns the extent to which one or more of the predictor variables (the Xs) in one's model can be predicted from other Xs in the model.

If Xs are "strongly" related, then

$$\left.\begin{array}{r} \hat{\beta}_j \text{ unreliable} \\ \\ \hat{\text{Var}}\ \hat{\beta}_j \text{ high} \\ \\ \text{model may not run} \end{array}\right\} \begin{array}{c} \text{collinearity} \\ \text{problem} \end{array}$$

If there are very strong relationships among some of the Xs, then the fitted model may yield unreliable regression coefficients for some predictors. In other words, coefficients may have high estimated variances, or perhaps the model may not even run. When this occurs, we say that the model has a *collinearity problem*.

Collinearity may involve more than two Xs

⇓

Simple approach : $r_{Xi,Xj}$ **not** sufficient

Because collinearity problems may involve relationships among more than two Xs, it is not sufficient to diagnose collinearity by simply looking at correlations among pairs of variables.

EXAMPLE

If $X3 \approx X1 - X2$,
could not detect this from
$r_{X3,X1}$, $r_{X3,X2}$, or $r_{X1,X2}$.

For example, if $X3$ was approximately equal to the difference between $X1$ and $X2$, this relationship could not be detected simply by looking at correlations between $X3$ and $X1$, $X3$ and $X2$, or $X1$ and $X2$.

EXAMPLE (continued)

COLLINEARITY DIAGNOSTIC TABLE

VDPs

		CNIs							
		99.8	30.8	12.2	8.6	5.0	2.1	1.3	1.0
	β_0	.325	.310	.020	.080	.002	.008	.250	.005
E	β_1	**.430**	**.440**	.021	.004	.001	.004	.000	.100
C_1	β_2	.114	.104	.046	.028	.695	.002	.000	.010
C_2	β_3	.030	**.819**	.018	.008	.003	.000	.101	.020
C_3	β_4	**.700**	.038	.024	.200	.014	.023	.000	.001
EC_1	β_5	.113	.102	.114	.022	.046	.200	.302	.101
EC_2	β_6	.018	**.963**	.007	.000	.000	.001	.000	.010
EC_3	β_7	**.930**	.010	.009	.006	.007	.008	.000	.030

diagnosing collinearity
1. Largest CNI "large" (e.g., >30)
2. At least two VDPs "large" (e.g., ≥ 0.5)

One popular way to diagnose collinearity uses a computer program or macro that produces a table (example shown at left) containing two kinds of information, *condition indices* (*CNI*s) and *variance decomposition proportions* (*VDPs*). (See Kleinbaum et al., Applied Regression and Other Multivariable Methods, 4th Edition, Chap. 14, 2008 for mathematical details about CNIs and VDPs)

Using such a table, a collinearity problem is diagnosed if the largest of the CNIs is considered large (e.g., >30) and at least two of the VDPs are large (e.g., ≥ 0.5).

		99.8	30.8	12.2	8.6	5.0	2.1	1.3	1.0
	β_0	.325	.310	.020	.080	.002	.008	.250	.005
E	β_1	**.430**	.440	.021	.004	.001	.004	.000	.100
C_1	β_2	.114	.104	.046	.028	.695	.002	.000	.010
C_2	β_3	.030	.819	.018	.008	.003	.000	.101	.020
C_3	β_4	**.700**	.038	.024	.200	.014	.023	.000	.001
EC_1	β_5	.113	.102	.114	.022	.046	.200	.302	.101
EC_2	β_6	.018	.963	.007	.000	.000	.001	.000	.010
EC_3	β_7	**.930**	.010	.009	.006	.007	.008	.000	.030

The diagnostic table we have illustrated indicates that there is at least one collinearity problem that involves the variables $\mathbf{E}$, $\mathbf{C_3}$ and $\mathbf{E} \times \mathbf{C_3}$ because the *largest CNI exceeds 30* and two of the *VDPs are as large as 0.5*.

Diagnosing collinearity conceptually
Computer software for nonlinear models

We now describe briefly how collinearity is diagnosed conceptually, and how this relates to available computer software for nonlinear models such as the logistic regression model.

Collinearity objective:

- Determine if fitted model is unreliable
$\Leftrightarrow$
- Determine whether $\text{Var}(\hat{\beta}_j)$ is "large enough"

The objective of collinearity diagnostics is to determine whether (linear) relationships among the predictor variables result in a fitted model that is "unreliable." This essentially translates to determining whether one or more of the estimated variances (or corresponding standard errors) of the $\hat{\beta}_j$ become "large enough" to indicate unreliability.

Estimated Variance–Covariance Matrix

$$\hat{\mathbf{V}} = \begin{bmatrix} \text{Var}(\hat{\beta}_1) & & \boxed{\text{Covariances}} & \\ & \text{Var}(\hat{\beta}_2) & & \\ & & \text{Var}(\hat{\beta}_3) & \\ \boxed{\text{Covariances}} & & & \text{Var}(\hat{\beta}_3) \end{bmatrix}$$

$= \mathbf{I}^{-1}$ for nonlinear models

The estimated variances are (diagonal) components of the estimated variance–covariance matrix ($\hat{\mathbf{V}}$) obtained for the fitted model. For non linear models in which ML estimation is used, the $\hat{\mathbf{V}}$ matrix is called *the inverse of the information matrix* ($\mathbf{I}^{-1}$), and is derived by taking the second derivatives of the likelihood function (L).

CNIs and VDPs derived from $\hat{\mathbf{V}}$

CNIs identify if collinearity exists

VDPs identify variables causing collinearity

The CNIs and VDPs previously introduced are in turn derived from the $\hat{\mathbf{V}}$ matrix. As illustrated earlier, the CNIs are used to identify whether or not a collinearity problem exists, and the VDPs are used to identify those variables that are the source of any collinearity problem. (Again, see Kleinbaum et al., 2008, Chapter 14, for a more mathematical description of CNIs, VDPs, and $\mathbf{I^{-1}}$.)

SAS, STATA, SPSS
do not compute CNIs and VDPs
for nonlinear models

But: SAS macro available

Unfortunately, popular computer packages such as SAS, STATA, and SPSS do not contain programs (e.g., SASs LOGISTIC procedure) that compute CNIs and VDPs for *nonlinear models*. However, a SAS macro (Zack et al.), developed at CDC and modified at Emory University's School of Public Health, allows computation of CNIs and VDPs for logistic and other nonlinear models (see Bibliography).

Application of macro later

We illustrate the use of this macro shortly.

Difficulties:

How large is large for CNIs and VDPs?
How to proceed if collinearity is found?

Nevertheless, there are difficulties in diagnosing collinearity. These include determining how "large is large" for both the CNIs and the VDPs, and how to proceed if a collinearity problem is found.

Collinearity cut-off recommendations

$$(\text{BKW}, 1981)\text{: CNI} \geq 30, \text{VDP} \geq 0.5$$

$$\Downarrow$$

Guidelines
for
linear regression models

The classic textbook on collinearity diagnostics (Belsey, Kuh, and Welch, 1981) recommends a cut-off of 30 for identifying a high CNI and a cut-off of 0.5 for identifying a high VDP. Nevertheless, these values were clearly described as "guidelines" rather than firm cut-points, and they were specified for linear models only.

Modifying guidelines for nonlinear models:

- open question (lower CNI cutpoint?)
- flexibility for **how high is high**

To what extent the guidelines (particularly for CNIs) should be modified (e.g., lowered) for nonlinear models remains an open question. Moreover, even for linear models, there is considerable flexibility in deciding how high is high.

We recommend (linear or logistic models):

- Require CNI >> 30
- Focus on VDPs for largest CNI
- Address largest CNI before other CNIs

For either linear or logistic models, we recommend that the largest CNI be "considerably" larger than 30 before deciding that one's model is unreliable, and then focusing on VDPs corresponding to the largest CNI before addressing any other CNI.

Sequential approach

⇓

Drop variable, refit model, readdress collinearity, continue until no collinearity

This viewpoint is essentially a *sequential* approach in that it recommends addressing the most likely collinearity problem before considering any additional collinearity problems.

Option 1 – (most popular) Correcting Collinearity:

EXAMPLE

Drop a variable from the model

Example: VDPs identify $X_1, X_2, X_1 \times X_2$

⇓

Drop $X_1 \times X_2$

Once a collinearity problem has been determined, the most popular option for correcting the problem is to drop one of the variables identified (by the VDPs) to be a source of the problem. If, for example, the VDPs identify two main effects and their product, the typical solution is to drop the product term from the model.

Dropping a collinear variable:

- Does not mean variable is nonsignificant
- Indicates dropped variable cannot be assessed with other collinear variables

Nevertheless, when such a term is dropped from the model, this does not mean that this term is nonsignificant, but rather that having such a term with other variables in the model makes the model unreliable. So, by dropping an interaction term in such a case, we indicate this interaction cannot be assessed, rather than it is nonsignificant.

Option 2 – Correction Collinearity: **Define a new (interpretable) variable:**

A second option for correcting collinearity is to define a new variable from the variables causing the problem, provided this new variable is (conceptually and/or clinically) interpretable.

- Does not make sense for product terms
- Can combine height and weight into BMI = height/weight^2

Combining collinear variables will rarely make sense if a product term is a source of the problem. However, if, for example, main effect variables such as height and weight were involved, then the "derived" variable BMI (= height/weight^2) might be used to replace both height and weight in the model.

EXAMPLE

MRSA example – Initial Model:

$$\begin{aligned}\text{Logit P}(\mathbf{X}) = \alpha &+ \beta_1 E_1 + \beta_2 E_2 + \gamma_1 V_1 \\ &+ \gamma_2 V_2 + \delta_{11} E_1 W_1 \\ &+ \delta_{12} E_1 W_2 + \delta_{21} E_2 W_1 \\ &+ \delta_{22} E_2 W_2 + \delta^* E_1 E_2,\end{aligned}$$

where

D = MRSA status $(0, 1)$,
E_1 = PREVHOSP
E_2 = PAMU, $V_1 = W_1$ = AGE,
$V_2 = W_2$ = GENDER

We now illustrate the use of collinearity diagnostics for the MRSA dataset we have described earlier. We consider the initial model shown at the left, which contains two *E*s, two *V*s, 4 *EW*s, and a single *EE*.

```
        COLLINEARITY DIAGNOSTICS FOR INITIAL MODEL
                          CNIs
          45.6 36.1 18.9 15.0 11.1  8.3  4.8  3.4  2.7  1.0
Intcept   0.39 0.56 0.04 0.00 0.01 0.02 0.00 0.00 0.00 0.00
Prehosp   0.95 0.00 0.01 0.03 0.00 0.00 0.00 0.00 0.00 0.00
AGE       0.32 0.50 0.00 0.12 0.02 0.01 0.01 0.01 0.00 0.00
gender    0.14 0.16 0.24 0.14 0.28 0.00 0.01 0.01 0.01 0.00
pamu      0.22 0.69 0.06 0.02 0.00 0.00 0.00 0.00 0.00 0.00
prhage    0.85 0.00 0.07 0.05 0.00 0.02 0.01 0.00 0.00 0.00
prhgen    0.35 0.01 0.33 0.26 0.01 0.00 0.03 0.00 0.01 0.00
prhpamu   0.03 0.03 0.63 0.22 0.01 0.07 0.00 0.01 0.00 0.00
pamage    0.26 0.53 0.08 0.00 0.08 0.03 0.00 0.00 0.00 0.00
pamgen    0.12 0.30 0.11 0.03 0.32 0.07 0.01 0.02 0.01 0.00
```

Using the collinearity macro introduced above, we obtain the (edited) collinearity diagnostic output shown at the left. From this table, we see that the highest CNI is 45.6, which is considerably higher than 30, and there are two VDPs greater than 0.5, corresponding to the variables PREVHOSP (VDP = 0.95) and the product term PREVHOSP × AGE (VDP = 0.85).

At least one collinearity problem, i.e., involves PREVHOSP and PREVHOSP × AGE ← Drop

Based on these results, we decide that there is at least one collinearity problem associated with the highest CNI and that this problem involves the two variables PREVHOSP and PREVHOSP × AGE.

```
      COLLINEARITY DIAGNOSTICS FOR REDUCED MODEL
                      CNIs          (w/o prhage)
          34.3 22.5  15.5  10.5   9.1   5.1   3.3   2.7   1.0
Intcept   0.73  0.22  0.00  0.01  0.03  0.00  0.00  0.00  0.00
prevhosp  0.01  0.61  0.32  0.01  0.00  0.05  0.00  0.00  0.00
AGE       0.74  0.03  0.12  0.03  0.07  0.00  0.02  0.00  0.00
gender    0.16  0.46  0.01  0.28  0.05  0.02  0.01  0.02  0.00
pamu      0.82  0.12  0.04  0.00  0.01  0.00  0.00  0.00  0.00
prhgen    0.01  0.82  0.04  0.01  0.06  0.04  0.00  0.01  0.00
prhpamu   0.04  0.13  0.74  0.00  0.07  0.01  0.01  0.00  0.00
pamage    0.75  0.00  0.03  0.12  0.09  0.01  0.00  0.01  0.00
pamgen    0.32  0.25  0.00  0.33  0.02  0.04  0.02  0.01  0.00
```

Proceeding sequentially, we would now drop the product term from the model and reassess collinearity for the resulting reduced model. The results are shown at the left. From this table, we see that the highest CNI is now 34.3, which is slightly higher than 30, and there are two VDPs greater than 0.5, corresponding to the variables AGE (VDP = 0.74) and the product term PAMU × AGE (VDP = 0.75).

Another possible collinearity problem (CNI = 34.3)
Two alternatives at this point:

- Stop further collinearity assessment
- Drop PAMU × AGE and continue

Since the highest CNI here (34.3) is only slightly above 30, we might decide that this value is not high enough to proceed further to assess collinearity. Alternatively, proceeding conservatively, we could drop the product PAMU × AGE and further assess collinearity.

EXAMPLE (continued)

```
   COLLINEARITY DIAGNOSTICS: 2nd REDUCED MODEL
                        CNIs   (w/o prhage and pamage)
          21.5  19.0  13.8   9.3  5.0    3.3   2.7   1.0
Intcept   0.20  0.65  0.12  0.01 0.00   0.01  0.00  0.00

prevhosp  0.59  0.08  0.28  0.00 0.046  0.00  0.00  0.00

AGE       0.02  0.24  0.37  0.32 0.01   0.03  0.01  0.00

gender    0.36  0.30  0.03  0.28 0.01   0.02  0.01  0.00

pamu      0.50  0.45  0.02  0.01 0.01   0.01  0.00  0.00

prhgen    0.84  0.00  0.06  0.02 0.06   0.00  0.01  0.00

prhpamu   0.18  0.30  0.44  0.05 0.00   0.02  0.00  0.00

pamgen    0.40  0.17  0.09  0.25 0.05   0.01  0.02  0.00
```

Reduced model after diagnosing collinearity:

$$\begin{aligned}\text{Logit P}(\mathbf{X}) = \alpha &+ \beta_1 E_1 + \beta_2 E_2 + \gamma_2 V_2\\ &+ \delta_{12} E_1 W_2 + \delta_{22} E_2 W_2\\ &+ \delta^* E_1 E_2\end{aligned}$$

(Note: $E_1 W_1$ and $E_2 W_1$ removed from initial model)

The collinearity diagnostics resulting when PAMU × AGE is dropped from the model are shown at the left. The largest CNI in this table is 21.5, which is much smaller than 30. Thus, we conclude that after we drop both PREVHOSP×AGE and PAMU×AGE, there are no more collinearity problems.

So, after assessing collinearity in our MRSA example, we have arrived at the reduced model shown at the left. This model then becomes a "revised" initial model from which we determine a final ("best") model using the hierarchical backward elimination (HBWE) strategy we have previously recommended.

V. Influential Observations

Are there any subjects in the dataset that "influence" study results?

⇓

Does removal of subject from the data result in "significant" change in $\hat{\beta}_j$ or $\hat{\text{O}}\text{R}$?

Another diagnostic issue concerns *influential observations:* those subjects (if any) in one's dataset that strongly "influence" the study results.

Technically, a subject is an influential observation if removal from the dataset results in a "significant" change in one or more of the estimated β_j (or ORs of interest in a logistic model).

Popular approach:
Measure extent of change in $\hat{\beta}_j$ when subject is dropped from the data:

Delta-beta ($\Delta\beta_j$)

A popular approach for identifying influential observations is to compute for each study subject, a measure of the change in one or more estimated regression coefficients when the subject is dropped from the data. For a given variable in the model, this measure is called a *Delta-beta*.

EXAMPLE

4 predictors: E, AGE, RACE, SEX ⇒
4 $\Delta\beta$ values for each subject (i)

$n = 100 \Rightarrow$ Compute 400 $\Delta\beta_j$ values

$\Delta\beta_{E,i}, \Delta\beta_{\text{AGE},i}, \Delta\beta_{\text{RACE},i}, \Delta\beta_{\text{SEX},i}$
i (subject) $= 1, 2, \ldots, 100$

$\Delta\beta_{E,i=\text{A}}$ is "large" or "significant":
⇓
removing subject A from analysis changes conclusions about effect of E

For example, a model containing four predictors, say E, AGE, RACE, and SEX, would produce four Delta-betas for each subject.

If the dataset contained 100 subjects, 400 Delta-betas would be computed, 4 for each subject.

If subject A, say, has a "large" or "significant" Delta-beta for the variable *E*, then one may conclude that removal of this subject from the analysis may change the conclusions drawn about the effect of *E*.

Summary measure for linear regression:

Cook's distance (CD)

combines $\Delta\beta_{j,i}$ **information** for all X_j predictors for subject i, e.g., a weighted average of the form

$$\text{CD}_\text{i} = \textstyle\sum_j w_j \Delta\beta_{j,i} \,/\, \sum_j w_j,$$
$$i = 1, 2, \ldots, n$$

Also, a summary measure that combines the Delta-beta information from all variables is typically computed. For linear regression, one such measure is called *Cook's distance*, which is a form of weighted average of the Delta-betas over all predictor variables X_j in one's model.

Logistic regression:

- **Cook's distance-type index**
- Uses approximation to change in logit values
- Similar to combining Delta-betas

For logistic regression, a similar measure (Pregibon, 1981) is used and often referred to as a *Cook's distance-type index*. This measure is derived using an approximation to the change in logit values when a subject is dropped from the data and, in essence, combines Delta-beta values using a logistic model.

Suggested alternative for logistic regression:

$$\text{CD}_i^* = \sum_j w_j \Delta(\exp[\beta])_{j,i} \,/\, \sum_j w_j$$

but not available in computer packages.

However, since the effect measure in logistic regression is typically an odds ratio, which exponentiates regression coefficients, a modified Cook's distance-type index that computes a weighted average of changes in exp[β], i.e., Δexp[β], might be preferable, but is not available in most computer packages.

More on influential observations including:

- How to use computer software?
- How to proceed when influential observations are identified?

We now briefly illustrate using the MRSA example how to use computer software to diagnose influential observations. We also discuss how to proceed with the analysis when influential observations are identified.

Computer Packages, e.g., SAS, STATA, SPSS produce their own version of influence diagnostics

Most computer software packages such as SAS, STATA, and SPSS allow the user to obtain influence diagnostics for logistic regression, although they each have their own version of the program code and the statistics produced.

SAS's LOGISTIC: **influence** and **iplots** options at end of model statement

For example, with SAS's LOGISTIC procedure, the user can specify two options: the "influence" option, and the "iplots" option after the model statement.

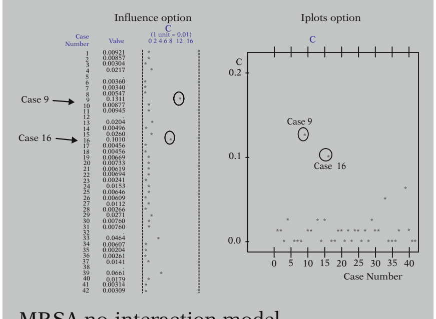
Large collection of regression diagnostic information produced, including **Δβ values** for each variable and **C measures** over all variables

Both these LOGISTIC options produce a large collection of regression diagnostics information for any fitted model. This includes Delta-beta measures for each variable in the model plus overall Cook's distance-type measures. Here, we focus on the latter, which we henceforth refer to as "C measures."

SAS's LOGISTIC: **C** and **Cbar** measures (similar but not identical)

Two slightly different C measures are produced by the influence and iplot options, a "C" and a "Cbar" measure (Pregibon, 1981). These measures typically yield similar, though not always identical, conclusions as to which subjects are "influential".

EXAMPLE



MRSA no-interaction model

$$\text{Logit } P(X) = \alpha + \beta_1\text{PREVHOSP} + \beta_2\text{PAMU} + \gamma_1\text{AGE} + \gamma_2\text{GENDER}$$

The **influence** option produces a figure that vertically (on the *Y*-axis) lists each subject and horizontally (on the *X*-axis) plots the value of the influence measure (C or Cbar). The **iplots** option, on the other hand, produces a figure that lists the subjects horizontally and plots the influence measure on the vertical axis.

The two figures on the left show the results for the influence measure C for the first 42 subjects in the MRSA data set for the no-interaction model shown below the figures. In both figures, subjects 9 and 16 appear to have C scores that are much higher than the other scores.

EXAMPLE (continued)

Results suggest that regression coefficients are "influenced" by these two subjects, e.g., drop subjects 9 and/or 16 from data

$\Downarrow$

Estimates of α, β_1, β_2, γ_1, γ_2 meaningfully change

Possibly influential subjects other than 9 and 16.

No Interaction Model w/o subjects 9 and 16

Param	DF	Estimate	Std Err	exp[coeff]
Intercept	1	−5.3830	0.8018	–
PREVHOSP	1	**1.6518**	0.4237	**5.217**
PAMU	1	**1.8762**	0.3809	**6.528**
AGE	1	0.0370	0.0095	1.038
GENDER	1	0.9214	0.3809	2.513

No Interaction Model full data

Param	DF	Estimate	Std Err	exp[coeff]
Intercept	1	−5.0583	0.7643	–
PREVHOSP	1	**1.4855**	0.4032	**4.417**
PAMU	1	**1.7819**	0.3707	**5.941**
AGE	1	0.0353	0.0092	1.036
GENDER	1	0.9329	0.3418	2.542

Should influential subjects be dropped from the data?

Answer: It depends!

- Incorrect data
- Incorrect model
- Legitimate and important data

These results indicate that if either or both of these subjects are dropped from the dataset, the collection of estimated regression coefficients in the fitted model would meaningfully change, which, in turn, could result in meaningfully different estimated ORs (e.g., $\exp[\hat{\beta}_1]$).

Since the above figures consider only 42 of a total of 289 subjects, there may be other influential subjects.

Without looking for other influential subjects, we show on the left the output obtained for the no-interaction model when subjects 9 and 16 are dropped from the dataset. Below this, we provide the output for the same model for the full dataset.

These results indicate that, particularly for the *E* variables PREVHOSP and PAMU, corresponding $\hat{\beta}_j$ and $\exp[\hat{\beta}_j]$ are somewhat different, although both sets of results indicate strong and statistically significant effects.

So, if we decide that some subjects (e.g., 9 and 16) are truly influential, what should we do? Drop them from the dataset?

The answer, once again, is it depends! A large influence statistic may be due to incorrect data on one or more subjects, but it can also be the result of an incorrect model, or even reflect the legitimate importance of (correct) data on a given subject.

Erroneous data

⇓

Correct if possible

Drop if not correctable

Certainly, if a subject's data is erroneous, it should be corrected if possible. If such an error is not clearly correctable, then the subject may be dropped from the analysis.

Correct data

⇓

Decision to delete up to researcher

Report and interpret in "discussion"

However, if the data on an influential subject is not erroneous, the researcher has to decide whether the subject should be dropped. For example, if the subject is much older than most subjects (i.e., an outlier), the researcher may have to decide whether the age range initially allowed needs to be modified. Instead of deleting such an individual, the researcher may wish to report and interpret the presence of influential subjects in the "discussion" of results.

Inappropriate model?

⇓

Difficult to decide if due to influential observation

It is typically difficult to determine whether a large influence statistic results from an inappropriate model. Since the initial model is rarely one's final (i.e., best) model, a final decision as to whether a given subject is influential should wait until one's final model is determined.

Summary:

Be careful about deleting.
Conservative approach:
drop subject only if
uncorrectable error

In summary, the researcher must be careful when considering whether or not to delete an observation. A very conservative approach is to only delete an observation if it is obviously in error and cannot be corrected.

VI. Multiple Testing

Modeling strategy ⇒ several statistical tests

$\Downarrow$

Potential for incorrect "overfitted" model, i.e., "too many" significant test results

The modeling strategy guidelines we have described when one's model contains either a single E (Chapters 6 and 7) or several Es (earlier in this chapter) all involve carrying out statistical significance testing for interaction terms as well as for E terms. Nevertheless, *performing several such tests on the same dataset may yield an incorrect "overfitted" final model* if "too many" test results are found to be significant.

"Too many":

variable(s) found significant, but H_0 true.

By "too many", we mean that the null hypothesis may actually be true for some significant test results, so that some "significant" variables (e.g., interaction terms) may remain in the final model even though the corresponding null hypotheses are true.

The multiple-testing problem:

Should we adjust for number of significance tests and, if so, how to adjust?

This raises the question as to whether or not we should adjust our modeling strategy to account for the number of statistical tests we perform and, if so, how should we carry out such adjustment?

Statistical principle:

number of significance tests increases

$\Downarrow$

$\alpha^* \longleftarrow \text{FWER}$

= Pr(Reject at least one H_{0i} | all H_{0i} true) increases

Note: α = **test-wise** error rate

= Pr(reject H_{0i} | H_0)

A well-established statistical inference principle is that the more statistical tests one performs, the more likely at least one of them will reject its null hypothesis even if all null hypotheses are true. The parameter α* shown at the left, is often called the *family-wise error rate* (FWER), whereas the significance level α for an individual test is called the *test-wise* error rate.

Formula: $\alpha^* = 1 - (1 - \alpha)^T$,

where T = number of independent tests of H_{0i}, $i = 1, \ldots, T$

Mathematically, the above principle can be expressed by the formula shown at the left.

EXAMPLE

$\alpha = 0.05 \Rightarrow$

T	α^*
1	0.05
5	0.23
10	0.40
20	0.64

For example, the table at the left shows that if $\alpha = 0.05$, and T ranges from 1 to 5 to 10 to 20, then α^* increases from 0.05 at $T = 1$ to 0.64 at $T = 20$.

Bonferroni approach:

To achieve $\alpha^* \leq \alpha_0$, set $\alpha = \alpha_0/T$

EXAMPLE

e.g., $\alpha_0 = 0.05$, $T = 10$:

$\alpha = 0.05/10 = 0.005$

$\Downarrow$

$\alpha^* = 1 - (1 - 0.005)^{10}$
$= 0.49 \leq \alpha_0 = 0.05$

A popular (Bonferroni) approach for insuring that α^* never exceeds a desired FWER of, say, α_0 is to require the significance level (α) for each test to be α_0/T. To illustrate, if $\alpha_0 = 0.05$ and $T = 10$, then $\alpha = 0.005$, and α^* calculates to 0.049, close to 0.05.

Problem with Bonferroni:

Over-adjusts: does not reject enough- low power (model may be underfitted)

A problem, however, with using the Bonferroni approach is that it "over-adjusts" by making it more difficult to reject any given H_{0i}; that is, its "power" to reject true alternative hypotheses is typically too low.

Bonferroni-type alternatives available to:

- Increase power
- Allow for nonindependent tests

Alternative formulae for adjusting for multiple-testing (e.g., Sidak, 1967; Holm, 1979; Hochberg, 1988) have been offered to provide increased power and to allow for nonindependent significance tests.

Another approach:

Replaces FWER with
False Discovery Rate (FDR) = T_0/T, where

T_0 = no. of tests incorrectly rejected, i.e., H_{0i} true
T = total no. of tests

Moreover, another adjustment approach (Benjamini and Hochberg, 1995) replaces the "overall" goal of adjustment from obtaining a desired "family-wise error rate" (FWER) to obtaining a desired "false discovery rate" (FDR), which is defined as the proportion of the number of significance tests that incorrectly reject the null (i.e., truly Type 1 errors).

Criticisms of multiple testing:

(1) Assuming **universal H$_0$:** all H_{0i} true unrealistic
(2) Paying a "penalty for peeking" reduces importance of specific tests of interest
(3) Where do you stop correcting for multiple-testing?

Nevertheless, there remains some controversy in the methodologic literature (Rothman, 1990) as to whether any attempt to correct for multiple-testing is even warranted. Criticisms of "adjustment" include (1) the assumption of a "universal" null hypothesis that all H_{0i} are non significant is unrealistic (2) paying a "penalty for peeking" (Light and Pillemer, 1984) reduces the importance of specific contrasts of interest; (3) where does the need for adjustment stop when considering all the tests that an individual researcher performs?

Multiple-testing literature: researcher knows in advance number of tests

Modeling strategy ("best model") problem: researcher does not know in advance number of tests

Finally, the literature on multiple-testing focuses on the situation in which the researcher knows in advance how many tests are to be performed. This is not the situation being addressed when carrying out a modeling strategy to determine a "best" model, since the number of tests, say for interaction terms, is only determined during the process of obtaining one's final model.

Bonferroni-type adjustment not possible when determining a "best model"

(Cannot specify *T* in advance)

Consequently, when determining a "best" model, a Bonferroni-type adjustment is not possible since the number of tests (*T*) to be performed cannot be specified in advance.

Ad hoc procedure:

Drop all variables in a nonsignificant chunk, e.g., all interaction terms

(Drawback: BW elimination may find significant effects overlooked by chunk test)

One approach for reducing the number of tests, nevertheless, is to use the results of non significant chunk tests to drop all the variables in the chunk, rather than continue with backward elimination (using more tests). However, note that the latter may detect significant (interaction) effects that might be overlooked when only using a chunk test.

Summary about multiple testing: No full-proof method available.

Thus, in summary, there is no full-proof method for adjusting for multiple-testing when determining a best model. It is up to the researcher to do anything, if at all.

VII. SUMMARY

This presentation is now complete.

Five issues on model strategy guidelines:

1. Modeling strategy when there are two or more exposure variables
2. Screening variables when modeling
3. Collinearity diagnostics
4. Multiple testing
5. Influential observations

We have described the five issues (shown at the left) on model strategy guidelines not covered in the previous two chapters on this topic.

Each of these issues represent important features of any regression analysis that typically require attention when determining a "best" model.

Issue 1: Several *E*s

$$\text{Logit P}(\mathbf{X}) = \alpha + \sum_{i=1}^{q} \beta_{\text{i}} E_{\text{i}} + \sum_{j=1}^{p_1} \gamma_{\text{j}} V_{\text{j}} + \sum_{i=1}^{q} \sum_{k=1}^{p_2} \delta_{ik} E_i W_k + \sum_{i=1}^{q} \sum_{\substack{i'=1 \\ i \neq i'}}^{q} \delta^{*}_{ii'} E_i E_{i'}$$

Regarding issue 1, we recommend that the initial model have the general form shown at the left. This model involves *E*s, *V*s, *EW*s, and *EE*s, so there are two types of interaction terms to consider.

<u>Modeling Strategy Summary: Several *E*s</u>

Step 1: Define initial model (above formula)
Step 2: Assess interaction: overall chunk test (?), then *EW*s and then (?) *EE*s
Step 3: Assess confounding and precision (*V*s) (prior to *EE*s?)
Step 4: Test for nonsignif *E*s if not components of significant *EE*s

We then recommend assessing interaction, first by deciding whether to do an overall chunk test, then testing for the *EW*s, after which a choice has to be made as to whether to test for the *EE* terms prior to or subsequent to assessing confounding and precision. The resulting model is then further assessed to see whether any of the *E* terms are nonsignificant.

Issue 2: Screening Variables
Method 0: Consider predictors one-at-a-time
Screen-out those X_i not significantly associated with *D*

Regarding issue 2, we described an approach (called *Method 0*) in which those variables that are not individually significantly associated with the (binary) outcome are screened-out (i.e., removed from one's initial model).

Does not consider confounding or interaction.
Questionable if model contains both *E*s and *C*s

Method 0 does not consider confounding and/or interaction for predictors treated one-at-a-time. Thus, Method 0 makes most sense when the model only involves *E*s but is questionable with both *E*s and *C*s being considered.

SUMMARY (*continued*)

Issue 3: Collinearity

Diagnose using CNIs and VDPs
Collinearity detected if:
Largest CNI is large (>30)
At least 2 VDPs are large (≥ 0.5)

For issue 3, we described how collinearity can be diagnosed from two kinds of information, *condition indices* (*CNI*s) and variance decomposition proportions (*VDP*s). A collinearity problem is indicated if the largest of the CNIs is considered large (e.g., >30) and at least two of the VDPs are large (e.g., ≥ 0.5).

Difficulties:
How large is large for CNIs and VDPs?
How to proceed if collinearity problem?

Nevertheless, difficulties remaining when assessing collinearity include how large is large for CNIs and VDPs, and how to proceed (e.g., sequentially?) once a problem is identified.

Issue 4: Influential Observations

Does removal of subject from the data result in "significant" change in $\hat{\beta}_j$ or $\widehat{\text{OR}}$?

Delta-beta ($\Delta\beta_j$): measures changes in specific β_j of interest
Cook's distance-type (C):
combines $\Delta\beta_j$ over all predictors (X_j)

Issue 4, concerning influential observations, is typically addressed using measures that determine the extent to which estimated regression coefficients are modified when one or more data points (i.e., subjects) are dropped from one's model. Measures that focus on such changes in specific regression coefficients of interest are called Delta-betas, whereas measures that combine changes over all regression coefficients in one's model are called Cook's distance-type measures.

Computer programs:
Provide plots of for each subject
Extreme plots indicate influential subjects

Computer programs for logistic regression models provide graphs/figures that plot such measures for each subject. Those subjects that show extreme plots are typically identified as being "influential."

Deleting influential observations:
Be careful!
Conservative approach: delete only if data in error and cannot be corrected

The researcher must be careful when considering whether or not to delete an observation. A conservative approach is to delete an observation only if it is obviously in error and cannot be corrected.

Issue 5: Multiple testing

The problem: should you adjust α when performing several tests?

Issue 5 (multiple testing) concerns whether or not the researcher should adjust the significance level used for significance tests to consider the number of such tests that are performed.

SUMMARY (*continued*)

Controversial issue:
Use Bonferroni-type adjustment
vs.
Do not do any adjustment

This is a controversial issue, in which various Bonferroni-type corrections have been recommended, but there are also conceptual arguments that recommend against any such adjustment.

When determining best model:
No well-established solution
No. of tests not known in advance

Nevertheless, when carrying out the process of finding a "best" model, there is no well-established method for such adjustment, since the number of tests actually performed cannot be known in advance.

We suggest that you review the material covered in this chapter by reading the detailed outline that follows. Then do the practice exercises and test.

In the next two chapters, we address two other regression diagnostic procedures: Goodness of fit tests and ROC curves.

Detailed Outline

I. Overview (page 244)

Focus: Five issues not considered in Chaps. 6 and 7

- Apply to any regression analysis but focus on binary logistic model
- Goal: determine "best" model
  1. Modeling strategy when there are two or more exposure variables
  2. Screening variables when modeling
  3. Collinearity diagnostics
  4. Influential observations
  5. Multiple testing

II. Modeling Strategy for Several Exposure Variables (pages 244–262)

A. Extend modeling strategy for (0,1) outcome, k exposures (Es), and p control variables (Cs)

B. Example with two Es: Cross-sectional study, Grady Hospital, Atlanta, GA, 297 adult patients Diagnosis: Staphylococcus aureus Infection

Question:

PREVHOSP, PAMU $\xrightarrow{?}$ MRSA,

controlling for AGE, GENDER

C. Modeling strategy summary: Several Es and Cs

Model: Logit P($\mathbf{X}$)

$$= \alpha + \sum_{i=1}^{q} \beta_i E_i + \sum_{j=1}^{p_1} \gamma_j V_j + \sum_{i=1}^{q} \sum_{k=1}^{p_2} \delta_{ik} E_i W_k$$

$$+ \sum_{i=1}^{q} \sum_{\substack{i'=1 \\ i \neq i'}}^{q} \delta^*_{ii'} E_i E_{i'}$$

Step 1: Define initial model (above formula)

Step 2: Assess interaction

Option A: Overall chunk test
+ Options B or C

Option B: Test EWs, then EEs

Option C: Test EWs, but assess Vs before EEs

Step 3: Assess confounding and precision (Vs)

Options A and B (continued): Vs after EWs and EEs

Options C (continued): Vs after EWs, but prior to EEs

Step 4: Test for nonsignif Es if not components of significant EEs

D. Modeling strategy: All *E*s, no *C*s

$$\text{Model: Logit P}(\mathbf{X}) = \alpha + \sum_{i=1}^{q} \beta_i E_i + \sum_{i=1}^{q} \sum_{\substack{i'=1 \\ i \neq i'}}^{q} \delta_{ii'}^{*} E_i E_{i'}$$

Step 1: Define initial model (above)

Step 2: Assess interaction involving *E*s.

Option A*: Overall chunk test for *EE*s, followed by backward elimination of *EE*s

Option B*: Skip chunk test for *EE*s; start with backward elimination of *EE*s

Skip previous Step 3

Step 4: Test for nonsignificant *E*s if not components of significant *EE*s

E. How causal diagrams can influence choice of initial model?

III. Screening Variables (pages 263–270)

A. Problem Focus: Model contains one *E*, and a large number of *C*s and $E \times C$s, **but** computer program does not run or fitted model unreliable ("large" p)

B. Screening: Exclude some C_j one-at-a-time; fit reduced model

C. **Method 0:** Consider predictors one-at-a-time; screen-out those X_i not significantly associated with the outcome (*D*)

D. Questions and Brief Answers about Method 0:

1. Any criticism? Yes: does not consider confounding or interaction involving *C*s
2. Depends on types of *X*s? Yes: use if only *E*s and no *C*s.
3. How large *k* compared to *n*? No good answer.
4. Other ways than Method 0? Yes: evaluate confounding and/or interaction for *C*s.
5. Collinearity and/or screening? Consider collinearity prior to and following screening.

E. Assessing Confounding and Interaction when Screening C variables.

Confounding: Compare Logit P($\mathbf{X}$) $= \alpha + \beta E$ with Logit P*($\mathbf{X}$) $= \alpha^* + \beta^* E + \gamma^* C$

Does $\widehat{\text{OR}}_{\text{DE}} = e^{\hat{\beta}} \neq \widehat{\text{OR}}_{\text{DE|C}} = e^{\hat{\beta}^*}$?

Interaction: Test H_0: $\delta = 0$ for the model Logit P($\mathbf{X}$) $= \alpha + \beta E + \gamma C + \delta EC$

F. How to proceed if several *E*s and several *C*s: It depends!

G. How to proceed if several *E*s and no *C*s: Use method 0.

IV. Collinearity (pages 270–275)

A. The Problem:

If predictors (Xs) are "strongly" related, then $\hat{\beta}_j$ unreliable, $\hat{\text{Var}}\hat{\beta}_j$ high, or model may not run.

B. Diagnosing Collinearity.

Use condition indices (CNIs) and variance decomposition proportions (VDPs).

Collinearity detected if: largest CNI is large (>30?) *and* at least 2 VDPs are large ($\geq$0.5?)

C. Collinearity for Logistic Regression

Requires computer macro (program not available in popular computer packages).

CNIs derived from inverse of Information Matrix ($\mathbf{I}^{-1}$)

D. Difficulties

How large is large for CNIs and VDPs? Guidelines provided are "soft."

How to proceed? We recommend sequential procedure: fix one collinearity problem at a time.

How to fix problem? Usual approach: drop one of the collinear variables; or, define new variable.

E. Example using MRSA data

V. Influential Observations (pages 275–279)

A. The Problem: Does removal of subject from the data result in "significant" change in $\hat{\beta}_j$ or $\widehat{\text{OR}}$?

B. Measures: Delta-betas ($\Delta\beta$s) and Cook's distance-type measures (Cs).

C. Computer packages: provide plots of $\Delta\beta$s and Cs for each subject.

D. What to do with influential observations:

Not easy to decide whether or not to drop subject from the data.

Conservative approach: drop subjects only if their data is incorrect and cannot be corrected.

VI. Multiple Testing (pages 280–282)

A. The Problem: should you adjust α when performing several tests?

B. Bonferroni approach: Use $\alpha = \alpha_0/T$, where α_0 = family-wise error rate (FWER) and T = number of tests.

C. Criticisms of Bonferroni approach: low power; based on unrealistic "universal H_0"; other.

D. Model building problem: number of tests (T) not known in advance; therefore, no foolproof approach.

VII. Summary (pages 283–285)

Practice Exercises

1. Consider the following logistic regression model in which all predictors are (0,1) variables:

$$\begin{aligned}\text{logit } P(\mathbf{X}) = {}& \alpha + \beta_1 E_1 + \beta_2 E_2 + \beta_3 E_3 + \gamma_1 C_1 + \gamma_2 C_2 + \gamma_3 C_3 \\ & + \gamma_{12} C_1 C_2 + \delta_{11} E_1 C_1 + \delta_{21} E_2 C_1 + \delta_{12} E_1 C_2 \\ & + \delta_{22} E_2 C_2 + \delta_{33} E_3 C_3 + \delta_{e13} E_1 E_3 + \delta_{e23} E_2 E_3 \\ & + \delta_{112} E_1 C_1 C_2 + \delta_{212} E_2 C_1 C_2\end{aligned}$$

For the above model, determine which of the following statements are **True** or **False.**
(i.e., **Circle T or F**)

T F a. The above model is hierarchically well-formulated.

T F b. Suppose the chunk test for the null hypothesis H_0: $\delta_{112} = \delta_{212} = 0$ is found to be significant and backward elimination involving these two three-factor product terms results in only $E_2C_1C_2$ remaining in the model. Then based on the "hierarchy principle," the final model must contain the variables C_1C_2, C_1, C_2, $E_2C_1C_2$, E_2C_1, E_2C_2, E_2E_3, and E_2.

T F c. Suppose the chunk test for the null hypothesis H_0: $\delta_{112} = \delta_{212} = 0$ is found to be significant and, as in the previous question, backward elimination involving these two three-factor product terms results in only $E_2C_1C_2$ remaining in the model. Then, based on the hierarchy principle and the hierarchical backward elimination approach, the only variables that remain as candidates for being dropped from the model at this point are E_1, E_3, E_1E_3, E_2E_3, E_1C_1, E_3C_3, and C_3.

T F d. Suppose that after the interaction assessment stage, the only terms remaining in the model are $E_2C_1C_2$, E_2C_1, E_2C_2, E_3C_3, C_1C_2, C_1, C_2, C_3, E_1, E_2, and E_3. Then, at this point, the odds ratio formula for comparing a person for whom $E_1 = E_2 = E_3 = 1$ to a person for whom $E_1 = E_2 = E_3 = 0$ is given by the expression

OR = $\exp[\beta_1 + \beta_2 + \beta_3 + \delta_{21}C_1 + \delta_{22}C_2 + \delta_{33}C_3 + \delta_{212}C_1C_2]$ where the coefficients in the formula are estimated from the reduced model obtained after interaction assessment.

T F e. Suppose that neither $E_1C_1C_2$ nor $E_2C_1C_2$ remains in the model after interaction assessment of these two three-factor products (but prior to interaction assessment of two-factor products). Suppose further that separate Wald (and corresponding likelihood ratio) tests for

H_0: $\delta_{11} = 0$, H_0: $\delta_{21} = 0$, and H_0: $\delta_{33} = 0$ are *nonsignificant* in the reduced model (without $E_1C_1C_2$ and $E_2C_1C_2$). Then, as the next step in interaction assessment, the model should be further reduced to

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha &+ \beta_1E_1 + \beta_2E_2 + \beta_3E_3 + \gamma_1C_1 + \gamma_2C_2\\ &+ \gamma_3C_3 + \gamma_{12}C_1C_2 + \delta_{12}E_1C_2\\ &+ \delta_{22}E_2C_2 + \delta_{e13}E_1E_3 + \delta_{e23}E_2E_3.\end{aligned}$$

prior to the assessment of confounding.

2. Suppose that after interaction assessment involving both EV_i and $E_i\ E_j$ terms in the initial model stated in question **1**, the following reduced model is obtained:

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha &+ \beta_1E_1 + \beta_2E_2 + \beta_3E_3 + \gamma_1C_1 + \gamma_2C_2\\ &+ \gamma_3C_3 + \gamma_{12}C_1C_2 + \delta_{11}E_1C_1 + \delta_{22}E_2C_2\end{aligned}$$

Suppose further that the assessment of confounding will *only* consider changes in the odds ratio that compares a person for whom $E_1 = E_2 = E_3 = 1$ to a person for whom $E_1 = E_2 = E_3 = 0$.

Based on the recommended guidelines for the assessment of confounding described in this chapter:

a. What is the formula for the *estimated* odds ratio in the gold standard model that should be used to assess confounding?

2 b. Assuming that you will need to consider tables of odds ratios that consider different subsets of potential confounders, *describe what a table of odds ratios would look like for the gold standard model* using the rectangle shown below for the (outside) borders of the table. In your answer, make sure to *state the formulae for the odds ratios* that will go into the different boxes in the table. *Hint.* You will need to draw horizontal and vertical lines to subdivide the rectangle and label the different row and column categories of the table, recognizing that the odds ratios being represented reflect the interaction effects that are present in the gold standard model.

c. Considering *only those variables that are candidates for being assessed as nonconfounders*,

   i. How many subsets of these variables need to be considered to address confounding?

   ii. List the variables that are contained in each of the above subsets.

d. Suppose that the following results are obtained when comparing tables of ORs for different subsets of V variables in the previously stated (question 2) reduced model obtained after interaction assessment:

Model #	Variables dropped from model	Table of ORs Within 10% of Gold Standard Table?
1	C_1	No
2	C_2	Yes
3	C_3	No
4a	C_1C_2	Yes
4b	C_1 and C_3	Yes
5	C_1 and C_1C_2	No
6	C_2 and C_1C_2	No
7	C_3 and C_1C_2	Yes
8	C_1 and C_2	Yes
9	C_1 and C_3	Yes
10	C_2 and C_3	Yes
11	C_1, C_2 and C_3	Yes
12	C_1, C_2 and C_1C_2	Yes
13	C_1, C_3 and C_1C_2	Yes
14	C_2, C_3 and C_1C_2	Yes
15	C_1, C_2, C_3, and C_1C_2	No
16	None	Yes (GS model)

Based on the above results, what models are eligible to be considered as final models after confounding assessment?

e. Based on your answer to part d, how would you address precision?

3. In addition to the variables E_1, E_2, E_3, C_1, C_2, C_3 considered in questions 1 and 2, there were 25 other variables recorded on study subjects that were identified from the literature review and conceptualization of the study as potential control variables. These variables were screened out by the investigators as not being necessary to include in the multivariable modeling analyses that were carried out.

Assume that screening was carried out by putting all 25 variables in a logistic regression model together with the variables E_1, E_2, E_3, C_1, C_2, and C_3 and then using a backward elimination to remove nonsignificant variables.

State or describe at least three issues/problems that would not have been addressed by the above screening approach.

4. Consider the following logistic regression model in which all predictor variables are (0, 1) variables:

$$\begin{aligned}\text{logit P}(\mathbf{X}) = {} & \alpha + \beta E + \gamma_1 C_1 + \gamma_2 C_2 + \gamma_3 C_3 + \gamma_{12} C_1 C_2 \\ & + \delta_1 EC_1 + \delta_2 EC_2 + \delta_3 EC_3 + \delta_{12} EC_1 C_2\end{aligned}$$

Determine which of the following statements are **True or False (Circle T or F).**

T F a. Suppose that using the collinearity macro described in the text results in the three highest condition indices (CNIs) being 77, 56, and 47. Then, using a sequential approach for diagnosing collinearity, there is at least 1 and possibly 3 collinearity problems associated with fitting this model.

T F b. Suppose that using the collinearity macro described in the text, the highest condition index (CNI) is found to be 77 and the only VDPs that are determined to be high for this CNI are associated with the variables E and C_3. Then it is reasonable to drop C_3 from the model, and recompute collinearity diagnostics to further reassess collinearity for the reduced model.

T F c. Suppose that using the collinearity macro described in the text, the three highest condition indices (CNIs) are 77, 56, and 47. For the highest CNI of 77, suppose there are exactly four VDPs other than the intercept that are larger than .5. Suppose also that the variables associated with these four VDPs are C_1, C_3, EC_3, and EC_1C_2. Then this collinearity problem can be addressed by dropping either EC_3 or EC_1C_2 from the model, but not C_1, C_2 or C_3, after which a reduced model is fit to see if there are additional collinearity problems.

T F d. Suppose that for the same situation as described in part c above, the collinearity problem is addressed by dropping EC_3 from the model, after which a reduced model without EC_3 is fit. Then the collinearity diagnostics obtained for the reduced model will indicate another collinearity problem that involves the variable EC_1C_2.

5. Suppose for the data analyzed using the model stated in question 4, it was desired to evaluate whether or not any subjects were influential observations.

   a. Assuming that the study objective was to assess the effect of E controlling for C_1, C_2, and C_3, how can you criticize the use of a Cooks-distance-type measure to identify influential subjects?

b. Suppose you identified five influential subjects in the dataset. How should you decide whether or not to drop any of these subjects from the data?

c. How can you criticize using the model stated in question 4 to identify influential subjects?

6. In attempting to determine a best model, starting with the model stated in question 4, suppose you wanted to use a Bonferroni-type adjustment to correct for multiple testing of interaction terms. What problem will you have in trying to determine the significance level for individual tests in order to achieve a family-wise error rate (FWER) not higher than 0.05?

Test

THE SCENARIO. A firm mattress is often believed to be beneficial for low back pain, although evidence supporting this recommendation is lacking. A randomized, double-blind, controlled, multicenter trial was carried out to determine the effect of mattress firmness on the clinical course of patients with chronic, nonspecific low back pain. A series of 313 adults with chronic, nonspecific low back pain, without referred pain, who complained of backache while lying in bed and upon rising were randomly assigned to four types of mattresses:
3 = firm, 2 = medium firm, 1 = medium soft, 0 = soft.
Clinical assessments were carried out at baseline and at 90 days. The endpoint (i.e., the health outcome variable) was improvement in pain while in bed and upon rising from bed between 0 and 90 days. The data set included the following variables:

D = improvement score for pain in bed from baseline to 90 days
(0 = no improvement or improvement in bed only, 1 = improvement upon rising only or improvement both in bed and upon rising)

F = firmness of mattress (3 = firm, 2 = medium firm, 1 = medium soft, 0 = soft)

BASE = type of base of bed (1 = firm, 0 = not firm)

POST = posture while sleeping at baseline (1 = supine or fetal, 0 = other)

PF = subjective perception of firmness of mattress (3 = firm, 2 = medium firm, 1 = medium soft, 0 = soft)

OCC = type of occupation (1 = sedentary, 0 = not sedentary)

AGE = age in years of subject

GEN = gender (0 = male, 1 = female)

Questions about how to analyze these data now follow:

1. In addition to the variables listed above, there were 12 other variables also listed in the dataset and identified from the literature review and conceptualization of the study as potential control variables. These variables were screened out by the investigators as not being necessary to include in the multivariable modeling analyses that were carried out.
   a. Assume that screening was carried out one variable at a time using tests of significance for the relationship between each potential control variable and the outcome variable, so that those potential control variables not found significantly associated with the outcome variable were not included in any modeling analysis.

      How can you criticize this approach to screening?
   b. Why was some kind of screening likely necessary for this analysis?

Suppose that the logistic regression treated the variables mattress type (**F**) and perceived mattress type (**PF**) as *ordinal* variables. Suppose also that the variable **BASE** is also considered to be an *exposure variable* (even though it was not involved in the randomization) in addition to the variable mattress type (**F**).

Suppose further that this model allows for two-way interactions (i.e., products of two variables) between mattress type (**F**) and each of the other independent variables (**POST, BASE, PF, OCC, AGE**, and **GEN**) and two-way interactions between **BASE** and each of the control variables **PF, POST, OCC, AGE**, and **GEN**.

2. State the logit formula for the logistic regression model just described. Make sure to consider both **F** and **BASE** as exposure variables.
3. For each of the following product terms, state whether the product term is an **EE** variable, an **EV** variable, or neither:

   F × POST

   F × BASE

   POST × BASE

   PF × POST

   BASE × PF

   (To answer this part, simply write **EE, EV, or neither** next to the variable regardless of whether the product term given is contained in the revised model described in question 2.)

4. Suppose that in carrying out interaction assessment for the model of question 2, *a chunk test for all two-way product terms is significant*. Suppose also that upon further interaction testing:

   a chunk test for two-way product terms involving **F** with **POST, OCC, PF, AGE,** and **GEN** has a *P*-value of 0.25 **and**

   a chunk test for two-way product terms involving **BASE** with **POST, OCC, PF, AGE,** and **GEN** has a *P*-value of 0.70.

Which of the following choices are "reasonable" as the next step in the assessment of interaction? **Circle as many choices as you consider reasonable.**

   a. All product terms except **F × BASE** should be dropped from the model.
   b. The two-way product terms involving **F** with **POST, OCC, PF, AGE,** and **GEN** should be dropped from the model and backward elimination should be carried out involving the two-way product terms involving **BASE** with **POST, OCC, PF, AGE,** and **GEN**.
   c. The two-way product terms involving **BASE** with **POST, OCC, PF, AGE,** and **GEN** should be dropped from the model and backward elimination should be carried out involving the two-way product terms involving **F** with **POST, OCC, PF, AGE,** and **GEN**.
   d. Carry out a backward elimination of all two-way product terms.

Note: In answering the above question, the word "reasonable" should be interpreted in the context of the hierarchical backward elimination strategy described in this text. Also, recall that **F** and **BASE** are exposure variables.

5. Describe how you would test for the significance of the two-way product terms involving **F with POST, OCC, PF, AGE,** and **GEN** using the model in question 2. In answering this question, make sure to state the null hypothesis in terms of model parameters, describe the formula for the test statistic, and give the distribution and degrees of freedom of the test statistic under the null hypothesis.

6. Suppose that at the end of the interaction assessment stage, it was determined that the variables **F × BASE, F × POST**, and **BASE × OCC** need to remain in the model as significant interaction effects. Based on the hierarchical backward elimination strategy described in Chap. 7, what **V** variables are eligible to be dropped

from the model as possible nonconfounders? Briefly explain your answer.

7. Based on the interaction assessment results described in question 6, is it appropriate to test the significance for the main effect of **POST** and/or **OCC**? Explain briefly.
8. a. **State the logit formula for the reduced model** obtained from the interaction results described in question 6.

   b. Based on your answer to 8 a, give a formula for the odds ratio that compares the odds for improvement of pain both in bed and upon rising to no improvement for a subject getting a firm mattress (**F** = 3) and a firm base (**BASE** = 1) to the corresponding odds for a subject getting a medium firm (**F** = 2) mattress and an infirm base (**BASE** = 0), controlling for **POST, PF, OCC, AGE**, and **GEN**.
9. Assume that the odds ratio formula obtained in question 8 represents the gold standard odds ratio for describing the relationship of mattress type (**F**) and mattress base (**BASE**) to pain improvement controlling for **POST, OCC, PF, AGE**, and **GEN**, i.e., *your only interest* is the **OR** comparing (**F** = 3, **BASE** = 1) with (**F** = 2, **BASE** = 0). One way to assess confounding among the variables eligible to be dropped as nonconfounders is to compare tables of odds ratios for each subset of possible confounders to the gold standard odds ratio.

   a. How many subsets of possible confounders (other than the set of possible confounders in the gold standard odds ratio) need to be considered?

   b. Describe what a table of odds ratios would look like for any of the subsets of possible confounders, i.e., draw such a table and specify what quantities go into the cells of the table.

   c. How would you use the tables of odds ratios described above to decide about confounding**?** In your answer, describe any difficulties involved.

   d. Suppose you decided that **PF** and **GEN** could be dropped from the model as nonconfounders, i.e., your reduced model now contains **F, BASE, POST, OCC, AGE, F × BASE, F × POST,** and **BASE × OCC:** Describe how you would determine whether precision was gained when dropping **PF** and **GEN** from the model.
10. Suppose that after all of the above analyses described in the previous questions, you realized that you

neglected to check for the possibility of **collinearity** in any of the models you considered so far.

a. Assuming that all the models you previously considered ran using the SAS's LOGISTIC procedure (i.e., they all produced output without any error or warning messages), why should you still be concerned about the possibility of collinearity?

Suppose, further, that you use SAS's LOGISTIC procedure to fit a model containing the independent variables **F, BASE, POST, OCC, PF, AGE, GEN**, all two-way product terms involving **F** with the control variables (**POST, OCC, PF, AGE, GEN**), and all two-way product terms involving **BASE** with these same control variables.

b. You now run the collinearity macro with the above logistic model, and you find that there are three condition indices with values 97, 75, and 62, with all other condition indices less than 25. How would you proceed "sequentially" to use this information to assess collinearity?

c. Suppose the condition index of 97 has "high VDP values" on the variables **F, BASE, PF**, and **F × BASE**. Does this result cause you difficulty in accepting your previous assessment of interaction in which you found that the variables **F × BASE, F × POST**, and **BASE × OCC** needed to remain in the model as significant interaction effects? Explain.

d. Based on the collinearity results in part 10 b and c, **which of the following choices is an appropriate next step?** *(Circle the "best" choice)*

   i. Drop the product term **F × BASE** from the polytomous logistic regression model and redo the hierarchical backward elimination strategy without further consideration of collinearity.

   ii. Determine whether the next highest condition index of 75 corresponds to high VDP loadings two or more predictors.

   iii. Drop the product term **F × BASE** from the logistic regression model, and apply collinearity diagnostics to the reduced model to determine if there is an additional collinearity problem.

   iv. Ignore the collinearity diagnostics results and use the model obtained from the hierarchical backward elimination strategy previously used.

11. a. Assuming that mattress type (**F**) and type of base (**BASE**) are the only two exposures of interest, with **PF, GEN, POST, OCC,** and **AGE** considered as control variables, briefly outline how you would

assess whether or not any subjects in the dataset are influential observations. Make sure to indicate whether you would prefer to use DeltaBeta measures or Cook's distance-type measures, both types of measures, or other measures.

b. Why would it be questionable to automatically drop from your dataset any subjects that you find to be influential observations?

12. Suppose in your analysis strategy for determining a "best" model, you want to reduce the number of statistical tests that you perform. What approach(es) can you use? Why are you not able to adequately carry out a Bonferroni-type of adjustment procedure?

Answers to Practice Exercises

1. a. True
   b. False: E_2E_3 not needed.
   c. False: E_1C_2 also a candidate
   d. True
   e. False: Incorrect use of backward elimination.
2. a. $\widehat{\text{OR}} = \exp[\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\delta}_{11}C_1 + \hat{\delta}_{22}C_2]$
   b.

	$C_2 = 1$	$C_2 = 0$
$C_1 = 1$	$\widehat{\text{OR}}_{11}$	$\widehat{\text{OR}}_{10}$
$C_1 = 0$	$\widehat{\text{OR}}_{01}$	$\widehat{\text{OR}}_{00}$

$$\widehat{\text{OR}}_{\text{ab}} = \exp[\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\delta}_{11}(C_1) + \hat{\delta}_{22}(C_2)]$$

   c. i. **4**
      ii. $\{C_3, C_1C_2\}$ $\{C_3\}$ $\{C_1C_2\}$ {Neither C_3 nor C_1C_2}
   d. Models 16, 7, 4a
   e. Obtain tables of confidence intervals for the odds ratios for each of the three models stated in part d. Choose as the best model either:
      i. the model with the narrowest confidence interval
      ii. the gold standard model (16) if all four models have approximately the same width.
      iii. the more parsimonious model
3. i. Confounding for individual Cs not addressed by statistical testing.
   ii. Interaction of individual Cs with each E_i not addressed
   iii. Interaction of E_i with E_j not addressed
   iv. Screening of Cs were not distinguished from screening of Es.

4. a. False: Can't tell until you check VDPs. Possible that all VDPs are not high (i.e., much less than 0.5)
   b. False: Model won't be HWF if C_3 is dropped.
   c. True
   d. False: May not be any remaining collinearity problem once EC_3 is dropped.
5. a. A Cook's distance-type measure combines the information from all estimated regression coefficients in one's model, whereas it would be preferable to consider either the $\Delta\beta$ or $\Delta\exp[\beta]$ for the E variable alone, since the E variable is the primary variable of interest.
   b. You should not automatically drop a subject from the dataset just because you have identified it as influential. A conservative approach is to drop only those subjects whose data are clearly in error and cannot be corrected.
   c. The model of question 4 may not be the best model, so that different conclusions might result about which subjects are influential if a different ("best") model were used instead.
6. The number of tests to be performed cannot be determined in advance of the modeling process, i.e., it is not clear what T will be for the individual significance level of $0.05/T$.

9 Assessing Goodness of Fit for Logistic Regression

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_9, © Springer Science+Business Media, LLC 2010

Introduction

Regression diagnostics are techniques for the detection and assessment of potential problems resulting from a fitted regression model that might either support, compromise, or negate the assumptions made about the regression model and/or the conclusions drawn from the analysis of one's data.

In this chapter, we focus on one important issue for evaluating binary logistic regression results, namely, goodness of fit (GOF) measurement. Although examination of data for potential problems, such as GOF, has always been considered a requirement of the analysis, the availability of computer software to efficiently perform the complex calculations required has contributed greatly to fine-tuning the diagnostic procedures and the conclusions drawn from them.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

Objectives

Upon completing this chapter, the learner should be able to:

1. Explain briefly what is meant by goodness of fit.
2. Define "perfect prediction."
3. Distinguish between "events–trials" format and "subject-specific" format for a dataset.
4. Define and illustrate the covariate patterns for a specific logistic model.
5. State or recognize the distinction between a fully parameterized and a saturated binary logistic model.
6. Given a specific binary logistic model, state or recognize the deviance formula for the model.
7. Explain briefly why the deviance statistic is not used for assessing goodness of fit when fitting a binary logistic regression model.
8. Given a printout of the results of a binary logistic regression:
   a. State or recognize the Hosmer–Lemeshow statistic
   b. Carry out a test of hypothesis for goodness of fit using the Hosmer–Lemeshow statistic

Presentation

I. Overview

Focus Does estimated logistic model predict observed outcomes in data?

This presentation describes methods for assessing the extent to which a logistic model estimated from a dataset predicts the observed outcomes in the dataset. The classical term for this topic is *goodness of fit (GOF)*.

- Considers a given model
- Does not consider comparing models

GOF is an issue that considers how well a given model, considered by itself, fits the data, rather than whether or not the model is more appropriate than another model.

Primary analysis goal:
Assess E–D relationship to derive "best" model

GOF goal:
Determine how well final ("best") model fits the data

In most epidemiologic analyses, the primary goal is to assess an exposure–disease relationship, so that we are usually more interested in deriving the "best" model for the relationship (which typically involves a strategy requiring the comparison of various models) than in using a GOF procedure. Nevertheless, once we have obtained a final (i.e., best") model, we would also like this model to fit the data well, thus justifying a GOF procedure.

Assume: Y is binary (0,1)
Units of analysis: individual subjects

GOF: Summary measure that compares

$$Y_i \text{ to } \hat{Y}_i, \quad \text{where}$$

$Y_i =$ observed response for subject i

$\hat{Y}_{\text{i}} = \hat{\text{P}}(\mathbf{X}_i) =$ predicted response for subject i

$i = 1, 2, \ldots, n$

Assuming that the outcome (Y) is binary, say coded as 0 or 1, and the unit of analysis is an individual subject, *GOF* typically requires a summary measure over all subjects that compares the observed outcome (Y_i) for subject i to the predicted outcome ($\hat{Y}_i$) for this subject obtained from the fitted model, i.e., $\hat{Y}_i = \hat{\text{P}}(\mathbf{X}_i)$.

Good fit: GOF measure "small" or "n.s."

Not sufficient evidence to conclude a bad fit

Lack of fit: Otherwise

If when determined collectively over all subjects, the GOF measure is "small" or "nonsignificant," we say the model has *good fit*, although, technically, we mean there is not sufficient evidence to conclude a bad fit. Otherwise, we say that the model has evidence of *lack of fit*.

Widely used GOF measure: **deviance**

But: deviance problematic for binary logistic regression

A widely used GOF measure for many mathematical models is called the *deviance*. However, as we describe later, for a binary logistic regression model, the use of the deviance for assessing GOF is problematic.

Popular alternative:
Hosmer–Lemeshow (HL) statistic

A popular alternative is the Hosmer–Lemeshow (*HL*) statistic. Both the deviance and the HL statistic will be defined and illustrated in this chapter.

II. Saturated vs. Fully Parameterized Models

GOF: Overall comparison between observed and predicted outcomes

As stated briefly in the previous overview section, a measure of goodness of fit (GOF) provides an overall comparison between observed (Y_i) and predicted values ($\hat{Y}_i$) of the outcome variable.

Perfect fit: $Y_i - \hat{Y}_i = 0$ for all i.

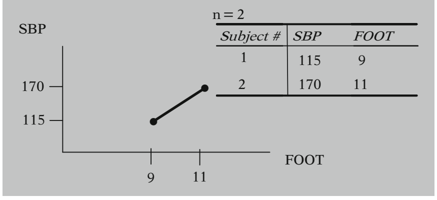
- Rarely happens
- Typically $0 < \hat{Y}_i < 1$ for most i

We say there is *perfect fit* if $Y_i - \hat{Y}_i = 0$ for all i. Although the mathematical characteristics of any logistic model require that the predicted value for any subject must lie between or including 0 or 1, it rarely happens that predicted values are either 0 or 1 for *all* subjects in a given dataset. In fact, the predicted values for most, if not all, subjects will lie above 0 and below 1.

Perfect fit: Not practical goal
Conceptual ideal
Saturated model
(a reference point)

Thus, achieving "perfect fit" is typically not a practical goal, but rather is a conceptual ideal when fitting a model to one's data. Nevertheless, since we typically want to use this "ideal" model as a reference point for assessing the fit of any specific model of interest, it is convenient to identify such a model as a *saturated model*.

EXAMPLE



n = 2

Subject #	*SBP*	*FOOT*
1	115	9
2	170	11

A trivial example of a saturated regression model is obtained if we have a dataset containing only $n = 2$ subjects, as shown on the left. Here, the outcome variable, SBP, is continuous, as is the (nonsense) predictor variable foot length (FOOT). A "perfect" straight line fits the data.

EXAMPLE (continued)

$\hat{\text{SBP}} = \hat{\beta}_0 + \hat{\beta}_1(\text{FOOT})$,

where $\hat{\beta}_0 = -132.5$ and $\hat{\beta}_1 = 27.5$

so $\hat{\text{SBP}} = -132.5 + 27.5(9) = 115$

and $\hat{\text{SBP}} = -132.5 + 27.5(11) = 170$

The linear regression model here involves only two parameters, β_0 and β_1, whose estimates yield predicted values equal to the two observed values of 115 and 170.

Linear model example:
$k + 1\ (= n) = 2$,
where $k + 1$ = # of parameters (including intercept)
and n = # of subjects

Thus, in this example, a saturated model is obtained when the number of model parameters ($k + 1 = 2$) is equal to the number of subjects in the dataset. (Note: k = # of variables in the model, and the "*1*" refers to the intercept parameter.)

Saturated Model (general):

$k + 1 = n$

where

$k + 1$ = # of parameters (including intercept)

n = sample size

More generally, the saturated model for a given dataset is defined as any model that contains as many parameters as the number of "observations" in the dataset, i.e., the sample size.

EXAMPLE

Observed Cohort Data

V = 1	E = 1	E = 0	
D = 1	6	4	10
D = 0	4	6	10

V = 0	E = 1	E = 0	
D = 1	3	7	10
D = 0	7	3	10

$\widehat{\text{OR}}_{V=1} = 2.250$ $\widehat{\text{OR}}_{V=0} = 0.184$

>1 <1

Very different

⇓

V is effect modifier of $\text{OR}_{E,D}$

To illustrate GOF assessment when using binary logistic regression, consider the following observed data from a cohort study on 40 subjects. The outcome variable is called D, there is one binary exposure variable (E), and there is one binary covariate (V).

These data indicate that V is an effect modifier of the E, D relationship, since the odds ratios of 2.250 and 0.184 are very different and are on opposite sides of the null value of 1.

Model 1: $\text{logit}\ P(\mathbf{X}) = \alpha + \beta E$

Model 2: $\text{logit}\ P(\mathbf{X}) = \alpha + \beta E + \gamma V$

Model 3: $\text{logit}\ P(\mathbf{X}) = \alpha + \beta E + \gamma V + \delta EV$

Three models that may be fit to these data are shown at the left. In model 1, E is the only predictor. In model 2, both E and V are predictors. Model 3 includes the product term $E \times V$ in addition to both E and V main effect terms.

$n = 40$ but $k + 1 = 2, 3$, or 4
i.e., $k + 1 < n$ in all 3 models
i.e., **no model is saturated**
(for predicting individual outcome)

Since the total sample size is 40, whereas each of these models contains 2, 3, and 4 parameters, respectively, *none* of these three models are saturated because $k + 1 < n$ for each model.

EXAMPLE (continued)

Key assumption:
Unit of analysis is the subject

Note, however, that concluding that "none of the models are saturated" is based on the following assumption: *the unit of analysis is the subject*.

Datalines listed by subject (e.g., Model 2)

Subject (i)	D	E	V
1	1	1	1
2	1	1	1
⋮	⋮	⋮	⋮
39	0	0	0
40	0	0	0

This assumption is equivalent to listing the dataset using 40 lines of data, one for each subject. For Model 2, therefore, each dataline would contain for a given subject, the values of the outcome (D) and the predictor variables (E and V) in the model.

If subjects are units of analysis, then Model 2 is not saturated
$(k + 1 = 3 < n = 40)$

When each subject is the unit of analysis, we can therefore claim that Model 2 is not saturated because the number of parameters (3) in the model is less than total units in the dataset ($n = 40$).

Alternative assumption:
Unit of analysis is a group
(subjects with same covariate pattern)

However, there is another way to view the dataset we have been considering: *the unit of analysis is a group of subjects*, all of whom have the same covariate pattern within a group.

Datalines Listed by Group (e.g., Model 2)

Group (g)	d_g	n_g	E	V
1	6	10	1	1
2	4	10	0	1
3	3	10	1	0
4	7	10	0	0

This assumption is equivalent to listing the dataset using only four lines of data, one for each covariate pattern. Each dataline would contain for a given group of subjects, the number of cases ($D = 1$) in each group (d_g), the number of subjects in each group (n_g), and the values of each predictor variable being modeled, as shown at the left for our dataset.

Events–trials format
(d_g) (n_g)

This type of data layout is called an *events–trials* format, where there are d_g events and n_g trials.

$n = \#$ of observations $= 4$
(Model 2)

Using events–trials format, we can argue that the number of observations (n) consists of the total number of datalines (4 in our example), rather than the number of subjects (40).

Goal of model prediction:
$\hat{p}_g = d_g/n_g$
(**group prediction** rather than individual prediction)

Here, the goal of model prediction no longer is to predict an individual's (0 or 1) outcome, but rather to predict the observed proportion of persons in a group (the unit of analysis) that has the outcome, i.e., $\hat{p}_g = d_g/n_g$.

Events–trials format:
"Group-saturated" provided $n = k + 1$
for n **covariate patterns**
(i.e., perfectly predicts $\hat{\mathbf{p}}_g$)

Thus, using events–trials format, we can declare a model to be "group-saturated" if for each covariate pattern listed in the dataset, the model perfectly predicts the observed proportion $\hat{p}_g$.

EXAMPLE (continued)

Model 3: logit P(**X**) $= \alpha + \beta E + \gamma V + \delta EV$

Largest possible model containing binary E and binary V

Note: $E^2 = E$ and $V^2 = V$

Let us now focus on Model 3. This model is the largest possible model that can be defined containing the two "basic" variables, E and V. Since E and V are both (0,1) variables, $E^2 = E$ and $V^2 = V$, so we cannot add to the model any higher order polynomials in E or V or any product terms other than $E \times V$.

Fully parameterized model: Contains maximum # of covariates defined from the main-effect covariates,

Model 3 is an example of *a fully parameterized model*, which contains the maximum number of covariates that can be defined from the main-effect covariates in the model.

No. of parameters $(k + 1) = G$ **covariate patterns**, where G = # of covariate patterns

Equivalently, the number of parameters in such a model must equal the number of *covariate patterns* (G) that can be defined from the covariates in the model.

Covariate patterns (i.e., subgroups): Distinct specifications of **X**

In general, for a given model with covariates $\mathbf{X} = (X_1, \ldots, X_k)$, the covariate patterns are defined by the distinct values of **X**.

EXAMPLE

Model 3: logit P(**X**) $= \alpha + \beta E + \gamma V + \delta EV$

4 covariate patterns

$\mathbf{X}_1$: $E = 1$, $V = 1$

$\mathbf{X}_2$: $E = 0$, $V = 1$

$\mathbf{X}_3$: $E = 1$, $V = 0$

$\mathbf{X}_4$: $E = 0$, $V = 0$

$k + 1 = 4 = G$ fully parameterized

Model 1: logit P(**X**) $= \alpha + \beta E$

2 covariate patterns

$\mathbf{X}_1$: $E = 1$

$\mathbf{X}_2$: $E = 0$

$k + 1 = 2 = G$ fully parameterized

Model 2: logit P(**X**) $= \alpha + \beta E + \gamma V$

4 covariate patterns (same as Model 3)

$k + 1 = 3 \neq G = 4$ *not* fully parameterized

For Model 3, which contains four parameters, there are four distinct covariate patterns, i.e., subgroups, that can be defined from the covariates in the model. These are shown at the left.

Model 1, which contains only binary E, is also fully parameterized, providing E is the only basic predictor of interest. No other variables defined from E, e.g., $E^2 = E$, can be added to model 1. Furthermore, Model 1 contains two parameters, which correspond to the two covariate patterns derived from E.

However, Model 2 is not fully parameterized, since it contains three parameters and four covariate patterns.

Assessing GOF:

Saturated $(k + 1 = n)$
vs.
fully parameterized $(k + 1 = G)$?

Thus, we see that a fully parameterized model has a nice property (i.e., $k + 1 = G$): it is the *largest possible model we can fit using the variables we want to allow into the model*. Such a model might alternatively be used to assess GOF rather than using the saturated model as the (gold standard) referent point.

Model A: $\mathbf{X} = (X_1, X_2, X_3)$ fully parameterized but

Model B: $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6)$ "better fit" than Model A

Note, however, even if a model A is fully parameterized, there may be a larger model B containing covariates not originally considered in model A that provides "better fit" than model A.

EXAMPLE

e.g.,

$$\text{LR test} \longrightarrow \begin{cases} \text{Model 1 : logit P}(\mathbf{X}) \\ \quad = \alpha + \beta E \\ \qquad \text{vs.} \\ \text{Model 3 : logit P}(\mathbf{X}) \\ \quad = \alpha + \beta E + \gamma V + \delta EV \end{cases}$$

$$\text{LR} = -2\ln\hat{L}_{\text{Model 1}} - (-2\ln\hat{L}_{\text{Model 3}})$$
$$= 55.051 - 51.355 = 3.696$$

$\sim \chi^2$(2 df) under H_0: $\gamma = \delta = 0$

($P > 0.10$) n.s.

For instance, although Model 1 is the largest model that can be defined when only binary E is considered, it is not the largest model that can be defined when binary V is also considered. Nevertheless, we can choose between Model 1 and Model 3 by performing a standard likelihood ratio (LR) test that compares the two models; the (nonsignificant) LR results ($P > 0.10$) are shown at the left.

Fitted Model 3:

logit $\hat{\text{P}}(\mathbf{X}) = \hat{\alpha} + \hat{\beta}E + \hat{\gamma}V + \hat{\delta}EV$,

where $\hat{\alpha} = 0.8473, \hat{\beta} = -1.6946,$
$\hat{\gamma} = -1.2528, \hat{\delta} = 2.5055$

Focusing now on Model 3, we show the fitted model at the left. Note that these same estimated parameters would be obtained whether we input the data by individual subjects (40 datalines) or by using an events–trials format (4 datalines). However, using events–trials format, we lose the ability to identify which subjects become cases.

Covariate pattern	**Obs. risk**	**Pred. risk**
$\mathbf{X}_1$: $E = 1$, $V = 1$	$\hat{p}_1 = 0.6$	$\hat{\text{P}}(\mathbf{X}_1) = 0.6$
$\mathbf{X}_2$: $E = 0$, $V = 1$	$\hat{p}_2 = 0.4$	$\hat{\text{P}}(\mathbf{X}_2) = 0.4$
$\mathbf{X}_3$: $E = 1$, $V = 0$	$\hat{p}_3 = 0.3$	$\hat{\text{P}}(\mathbf{X}_3) = 0.3$
$\mathbf{X}_4$: $E = 0$, $V = 0$	$\hat{p}_4 = 0.7$	$\hat{\text{P}}(\mathbf{X}_4) = 0.7$

The predicted risks obtained from the fitted model for each covariate pattern are also shown at the left, together with their corresponding observed risks. Notice that these predicted risks are equal to their corresponding observed proportions computed from the observed (stratified) data.

$E = 1, V = 1$: $\hat{\text{P}}(\text{X}_1) = 0.6 \neq 0$ or 1

$E = 0, V = 1$: $\hat{\text{P}}(\text{X}_2) = 0.4 \neq 0$ or 1

Model 3: No perfect fit – $E = 1$: some $D = 1$, some $D = 0$
$E = 0$: some $D = 1$, some $D = 0$

Nevertheless, none of these predicted risks are either 0 or 1, so the fitted model does not perfectly predict each subject's observed outcome, which is either 0 or 1. This is not surprising, since some exposed subjects develop the disease and some do not.

Model 3. Fully parameterized model

$$\Downarrow$$

# of expected cases
= # of observed cases for
each covariate pattern

However, we will show below that because Model 3 is a fully parameterized model, it perfectly predicts the number of cases actually observed for each covariate pattern. That is, the expected number of cases for each pattern, based on the fitted model, equals the observed number of cases for each pattern.

Covariate pattern $\mathbf{X}_g$: n_g subjects
$\mathbf{X}_g$ = values of **X** in group g
d_g = observed cases in group g (binomial)
$\hat{P}(\mathbf{X}_g)$: predicted risk in group g
$\hat{d}_g = n_g\hat{P}(\mathbf{X}_g)$: expected cases in group g

More specifically, suppose n_g subjects have covariate pattern $\mathbf{X}_g$, and d_g denotes the *observed number of cases* in group g. Then, since d_g has the binomial distribution, the *expected number of cases* in group g is $\hat{d}_g = n_g\hat{P}(\mathbf{X}_g)$, where $\hat{P}(\mathbf{X}_g)$ is the predicted risk for any subject in that group.

EXAMPLE

X: EV	Exp. Cases	Obs. Cases
$\mathbf{X}_1$:11	$\hat{d}_1 = 10(0.6) = 6$	$d_1 = 6$
$\mathbf{X}_2$:01	$\hat{d}_2 = 10(0.4) = 4$	$d_2 = 4$
$\mathbf{X}_3$:10	$\hat{d}_3 = 10(0.3) = 3$	$d_3 = 3$
$\mathbf{X}_4$:00	$\hat{d}_4 = 10(0.7) = 7$	$d_4 = 7$

Model 3.	Perfect Prediction?
$Y_i - \hat{Y}_i \neq 0$ for all *i* (not saturated)	**Individuals: No**
$d_g - \hat{d}_g = 0$ for all g (fully parameterized)	**Groups: Yes** (Patterns)

Model 3 is "group-saturated": perfectly group outcomes

Thus, for Model 3, we expect $\hat{d}_1 = 10(0.6) = 6$ cases among subjects with $E = 1$ and $V = 1$, $\hat{d}_2 = 10(0.4) = 4$ cases among subjects with $E = 0$ and $V = 1$, and so on for the other two covariate patterns. The corresponding observed number of subjects are also shown at the left. Notice that the corresponding observed and expected cases are equal for Model 3.

So, even though Model 3 does not provide "perfect prediction" in terms of *individual* outcomes, it does provide "perfect prediction" in terms of *group* outcomes.

In other words, although $Y_i - \hat{Y}_i \neq 0$ for all subjects, $d_g - \hat{d}_g = 0$ for all covariate patterns.

Another way of saying this is that Model 3 is "group-saturated" in the sense that Model 3 perfectly predicts the group outcomes corresponding to the distinct covariate patterns.

Two GOF approaches:
Compare fitted model to:

1. Saturated model: Provides perfect **individual prediction**
2. Fully parameterized model: Provides perfect **group prediction** (based on covariate patterns)

Thus, we see that an alternative gold standard model for assessing GOF is a fully parameterized (group-saturated) model containing the covariates of interest rather than a (subject-specific) saturated model that can rarely if ever be achieved using these covariates.

Classical GOF approach:

Saturated model gives perfect fit for individual subjects

Why?

$Y_i = 0$ or 1 only possible outcomes for subject i

However, problematic for logistic regression

The traditional (i.e., "classical") GOF approach considers the saturated model as the ideal for "perfect fit." This makes sense when the units of analysis are *individual subjects*, since their actual observed outcomes are 0 or 1, rather than some value in between. However, as we will explain further (later below), use of the saturated model to assess GOF for logistic regression is problematic.

Saturated model: $k + 1 = n$

EXAMPLE

Previous example ($n = 40$, 4 covariate patterns):

Model 4 (SS saturated model)

$$\text{logit P}(\mathbf{X}) = \omega_1 Z_1 + \omega_2 Z_2 + \omega_3 Z_3 + \cdots + \omega_{40} Z_{40}$$

$$Z_i = \begin{cases} 1 \text{ if subject } i; \quad i = 1, 2, \ldots, 40 \\ 0 \text{ otherwise} \end{cases}$$

$$L_{SS} = \prod_{i=1}^{40} \text{P}(\mathbf{X}_i)^{Y_i} (1 - \text{P}(\mathbf{X}_i))^{1-Y_i}$$

where $\mathbf{X}_i$ denotes the values of $\mathbf{X}$ for subject i

Recall that we originally defined the saturated model as that model for which the number of parameters ($k + 1$) equals the sample size (n). For our example involving four covariate patterns for 40 subjects, the *subject-specific* (SS) saturated model is shown at the left. This model does not have an intercept term, but does contain 40 parameters, as defined by the ω_i. The Z_i are dummy variables that distinguish the 40 subjects.

The likelihood function for this (SS) saturated model is shown at the left. In this formula, Y_i denotes the observed value (either 0 or 1) for the ith individual in the dataset.

Subject $i : Z_i = 1$, other $Zs = 0$

$$\Downarrow$$

logit $\text{P}(\mathbf{X}_i) = \omega_i$ and

$\text{P}(\mathbf{X}_i) = 1/[1 + \exp(-\omega_i)]$

Note that for subject i, $Z_i = 1$ and $Z_k = 0$ for $k \neq i$, so $\text{P}(\mathbf{X}_i)$ can be written in terms of the regression coefficient ω_i that involves only that one subject.

Saturated model

$$\Downarrow$$

$$\hat{Y}_i \overset{\text{def}}{\equiv} \hat{\text{P}}(\mathbf{X}_i) = Y_i, \; i = 1, 2, \ldots, n$$

Furthermore, since the saturated model perfectly fits the data, it follows that the maximum likelihood (ML) estimate $\hat{Y}_i$, which equals $\hat{\text{P}}(\mathbf{X}_i)$ by definition, must be equal to the observed Y_i for each subject i.

$$\hat{L}_{\text{SS max}} = \prod_{i=1}^{40} Y_i^{Y_i}(1 - Y_i)^{1-Y_i}$$

It follows that the formula for the ML value of the SS saturated model ($\hat{L}_{\text{SS max}}$) involves substituting Y_i for P($\mathbf{X}_i$) in the above formula for the likelihood, as shown at the left.

For any binary logistic model:

$$Y_i = 1 : Y_i^{Y_i}(1 - Y_i)^{1-Y_i} = 1^1(1 - 1)^{1-1} = 1$$

$$Y_i = 0 : Y_i^{Y_i}(1 - Y_i)^{1-Y_i} = 0^0(1 - 0)^{1-0} = 1$$

From simple algebra, it also follows that the expression $Y_i^{Y_i}(1 - Y_i)^{1-Y_i}$ will always be equal to one when Y_i is a (0,1) variable.

$\hat{L}_{\text{SS max}} \equiv 1$ always

Important implications for GOF

Consequently, the maximized likelihood will always equal 1. This result has important implications when attempting to assess GOF using a saturated model as the gold standard for comparison with one's current model.

III. The Deviance Statistic

Deviance:

$$\text{Dev}(\hat{\boldsymbol{\beta}}) = -2\ln(\hat{L}_c / \hat{L}_{\max})$$

As mentioned at the beginning of this chapter, a widely used measure of GOF is the *deviance*. The general formula for the deviance (for any regression model) is shown at the left. In this formula, $\hat{\boldsymbol{\beta}}$ denotes the collection of estimated regression coefficients in the current model being evaluated, $\hat{L}_c$ denotes the maximized likelihood for the current model, and $\hat{L}_{\max}$ denotes the maximized likelihood for the saturated model.

$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$

$\hat{L}_c$ = ML for current model

$\hat{L}_{\max}$ = ML for saturated model

(Note: If subjects are the unit of analysis, $\hat{L}_{\max} \equiv \hat{L}_{\text{SS max}}$)

$\hat{L}_c$ closer to $\hat{L}_{\max}$

$\Downarrow$

better fit (smaller deviance)

Thus, the deviance contrasts the likelihood of the current model with the likelihood of the model that perfectly predicts the observed outcomes. The closer are these two likelihoods, the better the fit (and the smaller the deviance).

perfect fit

$\hat{L}_c = \hat{L}_{\max} \Rightarrow -2\ln(\hat{L}_c / \hat{L}_{\max}) = -2\ln(1) = 0$

$\hat{L}_c << \hat{L}_{\max} \Rightarrow \hat{L}_c / \hat{L}_{\max}$ small fraction

$\Rightarrow \ln(\hat{L}_c / \hat{L}_{\max})$ large,negative

poor fit

$\Rightarrow -2\ln(\hat{L}_c / \hat{L}_{\max})$ large,positive

In particular, if $\hat{L}_c = \hat{L}_{\max}$, then the deviance is 0, its minimum value. In contrast, if $\hat{L}_c$ is much smaller than $\hat{L}_{\max}$, then the ratio $\hat{L}_c/\hat{L}_{\max}$ is a small fraction, so that the logarithm of the ratio is a large negative number and −2 times this large negative number will be a large positive number.

Properties of deviance similar to properties of χ^2 statistic

These properties of the deviance, i.e., its values range from zero to larger and larger positive numbers, correspond to the properties of a chi-square statistic as used in a likelihood ratio test.

Common to test for GOF by comparing deviance with χ^2_{n-k-1} value. (questionably legitimate)

In fact, when using the deviance to test for GOF, it is common, though not strictly legitimate, to compare the deviance to chi-square values with $n-k-1$ degrees of freedom when the current model contains $k+1$ parameters.

GOF H_0: model fits
H_A: model does not fit

Deviance "significant" ⇒ poor fit

The GOF null hypothesis is that the "model fits," and the alternative hypothesis is that the "model does not fit." Thus, if the deviance is "significantly" large, the model is considered to have poor fit.

$$\begin{aligned}\text{Dev}(\hat{\boldsymbol{\beta}}) &= -2\ln(\hat{L}_{\text{c}}/\hat{L}_{\max}) \\ &= -2\ln\hat{L}_{\text{c}} - (-2\ln\hat{L}_{\max})\end{aligned}$$

Recall

$$\text{LR} = -2\ln\hat{L}_{\text{R}} - (-2\ln\hat{L}_{\text{F}}) \sim \chi^2$$

R = reduced model
F = full model

Note that the *deviance statistic is, by definition, a likelihood ratio (LR) statistic for comparing one's current model to the saturated model*. Thus, the use of the chi-square distribution to test for the significance of the deviance appears justified because the LR statistic has an approximate chi-square distribution under H_0 when comparing full vs. reduced (nonsaturated) models that are fitted using ML estimation.

$$\begin{aligned}&\text{Dev}_{\text{R}}(\hat{\boldsymbol{\beta}}) - \text{Dev}_{\text{F}}(\hat{\boldsymbol{\beta}}) \\ &= [-2\ln(\hat{L}_{\text{R}}/\hat{L}_{\max})] - [-2\ln(\hat{L}_{\text{F}}/\hat{L}_{\max})] \\ &= [-2\ln\hat{L}_{\text{R}} - \cancel{(-2\ln\hat{L}_{\max})}] - [-2\ln\hat{L}_{\text{F}} - \cancel{(-2\ln\hat{L}_{\max})}] \\ &= -2\ln\hat{L}_{\text{R}} - (-2\ln\hat{L}_{\text{F}}) \equiv \text{LR}\end{aligned}$$

In particular, it can be shown from simple algebra that the *LR test for comparing two hierarchical nonsaturated regression models is equivalent to the difference in deviances* between the two models. This follows, as shown at the left, because the maximized likelihood for the saturated model drops out of the difference in deviance scores.

Nevertheless for the **logistic model:** χ^2 approximation of deviance statistic is **questionable** (see below)

Nevertheless, as we shall soon explain, when using the deviance statistic to assess GOF for a single (nonsaturated) *logistic model*, the chi-square approximation for the LR test is questionable.

Alternative (Events–Trials) Deviance formula: for the logistic model:

G covariate patterns

$\mathbf{X}_g = (\mathrm{X}_{g1}, \mathrm{X}_{g2}, \ldots, \mathrm{X}_{gp}), g = 1, 2, \ldots, G$

$\hat{d}_{\mathrm{g}} = n_{\mathrm{g}}\hat{\mathrm{P}}(\mathbf{X}_g)$ = expected cases

d_{g} = observed cases

First, we present an alternative formula for the deviance in a logistic model that considers the covariate patterns defined by one's current model. We assume that this model contains G covariate patterns $\mathbf{X}_g = (X_{g1}, X_{g2}, \ldots, X_{gp})$, with n_{g} subjects having pattern g. As defined earlier, $\hat{d}_{\mathrm{g}} = n_{\mathrm{g}}\hat{\mathrm{P}}(\mathbf{X}_g)$ denotes the expected cases, where $\hat{\mathrm{P}}(\mathbf{X}_g)$ is the predicted risk for $\mathbf{X} = \mathbf{X}_g$, and d_{g} denotes the observed number of cases in subgroup g.

$$\begin{aligned}
&\mathrm{Dev}_{\mathrm{ET}}(\hat{\boldsymbol{\beta}}) \\
&\quad = -2\ln\hat{L}_{\mathrm{c,ET}} - (-2\ln\hat{L}_{\mathrm{max,ET}}) \\
&\quad = -2\sum_{g=1}^{\mathrm{G}}\left[d_{\mathrm{g}}\ln\left(\frac{d_{\mathrm{g}}}{\hat{d}_{\mathrm{g}}}\right)\right. \\
&\qquad\quad \left. + (n_{\mathrm{g}} - d_{\mathrm{g}})\ln\left(\frac{n_{\mathrm{g}} - d_{\mathrm{g}}}{n_{\mathrm{g}} - \hat{d}_{\mathrm{g}}}\right)\right],
\end{aligned}$$

where $-2\ln\hat{L}_{\mathrm{c,ET}}$ and $-2\ln\hat{L}_{\mathrm{max,ET}}$ are defined using events–trials (ET) format

The alternative deviance formula is shown here at the left. This formula corresponds to the dataset listed in events–trials format, where there are *G* datalines, d_{g} and n_{g} denote the number of events and number of trials, respectively, on the gth dataline.

EXAMPLE

Model 3: logit $\mathrm{P}(X) = \alpha + \beta E + \gamma V + \delta EV$

X: EV	n_{g}	Exp. Cases	Obs. Cases
$\mathbf{X}_1$: 11	10	$\hat{d}_1 = 6$	$d_1 = 6$
$\mathbf{X}_2$: 01	10	$\hat{d}_2 = 4$	$d_2 = 4$
$\mathbf{X}_3$: 10	10	$\hat{d}_3 = 3$	$d_3 = 3$
$\mathbf{X}_4$: 00	10	$\hat{d}_4 = 7$	$d_4 = 7$

$\mathrm{Dev}_{\mathbf{ET}}(\hat{\boldsymbol{\beta}})$ for Model 3:

$$\begin{aligned}
&= -2\left[6\ln\left(\frac{6}{6}\right) + 4\ln\left(\frac{4}{4}\right)\right] \\
&\quad -2\left[4\ln\left(\frac{4}{4}\right) + 6\ln\left(\frac{6}{6}\right)\right] \\
&\quad -2\left[3\ln\left(\frac{3}{3}\right) + 7\ln\left(\frac{7}{7}\right)\right] \\
&\quad -2\left[7\ln\left(\frac{7}{7}\right) + 3\ln\left(\frac{3}{3}\right)\right] \\
&= \mathbf{0}
\end{aligned}$$

Recall that for the data set on $n = 40$ subjects described above, Model 3 has $G = 4$ covariate patterns. For each pattern, the corresponding values for n_{g}, $\hat{d}_{\mathrm{g}}$, and d_{g} are shown at the left.

Substituting the values in the table into the alternative (events–trials) deviance formula, we find that the resulting deviance equals zero.

Deviance formula $\mathrm{Dev}_{\mathbf{ET}}(\hat{\boldsymbol{\beta}})$ uses events–trials format

⇓

Units of analysis are groups (not subjects)

How can we explain this result, since we know that the fully parameterized Model 3 is not saturated in terms of perfectly predicting the 0 or 1 outcome for each of the 40 subjects in the dataset? The answer is that since the ET deviance formula corresponds to an events–trials format, the units of analysis being considered by the formula are the four groups of covariate patterns rather than the 40 subjects.

EXAMPLE

"group-saturated":

# of parameters = # of groups

↑

sample size = 4

Consequently, Model 3 is "group-saturated" in that the number of parameters in this model is equal to the total number of groups being considered. That is, since the units of analysis are the four groups, the sample size corresponding to the alternative deviance formula is 4, rather than 40.

Model 3: $\text{logit P(X)} = \alpha + \beta E + \gamma V + \delta EV$

vs.

Model 4 (perfectly predicts 0 or 1 outcome):

$$\text{logit P}(\mathbf{X}) = \omega_1 Z_1 + \omega_2 Z_2 + \omega_3 Z_3 + \cdots + \omega_{40} Z_{40}$$

So, then, how can we compare Model 3 to the saturated model we previously defined (Model 4) that perfectly predicts each subject's 0 or 1 outcome?

Subject-specific deviance formula:

$$\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) = -2\sum_{i=1}^{n}\left[Y_i \ln\left(\frac{Y_i}{\hat{Y}_i}\right) + (1-Y_i)\ln\left(\frac{1-Y_i}{1-\hat{Y}_i}\right)\right]$$

Y_i = observed (0, 1) response for subject i

$\hat{Y}_i$ = predicted probability for the subject i

$\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) \neq \text{Dev}_{\text{ET}}(\hat{\boldsymbol{\beta}})$ unless $G = n$

$\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) > 0$ since $G << n$ for Model 3

This requires a second alternative formula for the deviance, as shown at the left. Here, the summation (i) covers all subjects, not all groups, and Y_i and $\hat{Y}_i$ denote the observed and predicted response for the ith subject rather than the gth covariate pattern/group.

This alternative (subject-specific) formula is not identical to the events–trials formula given above unless $G = n$. Moreover, since $G << n$ for Model 3, the SS deviance will be nonzero for this model.

Equivalent formula for $\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}})$:
(from calculus and algebra)

$$\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) = -2\sum_{i=1}^{n}\left[\hat{\text{P}}(\mathbf{X_i})\ln\left(\frac{\hat{\text{P}}(\mathbf{X_i})}{1-\hat{\text{P}}(\mathbf{X_i})}\right) + \ln(1-\hat{\text{P}}(\mathbf{X_i}))\right]$$

We now show how to compute the subject-specific (SS) deviance formula for Model 3. However, we first provide an equivalent SS deviance formula that can be derived from both algebra and some calculus (*see this chapter's appendix for a proof*).

EXAMPLE

Covariate pattern	Obs. risk	Pred. risk
$\mathbf{X}_1$: $E = 1$, $V = 1$	$\hat{p}_1 = 0.6$	$\hat{\text{P}}(\mathbf{X}_1) = 0.6$
$\mathbf{X}_2$: $E = 0$, $V = 1$	$\hat{P}_2 = 0.4$	$\hat{\text{P}}(\mathbf{X}_2) = 0.4$
$\mathbf{X}_3$: $E = 1$, $V = 0$	$\hat{p}_3 = 0.3$	$\hat{\text{P}}(\mathbf{X}_3) = 0.3$
$\mathbf{X}_4$: $E = 0$, $V = 0$	$\hat{p}_4 = 0.7$	$\hat{\text{P}}(\mathbf{X}_4) = 0.7$

To compute this formula for Model 3, we need to provide values of $\hat{\text{P}}(\mathbf{X}_i)$ for each of the 40 subjects in the data set. Nevertheless, this calculation can be simplified since there are only four distinct values of $\hat{P}(\mathbf{X}_\text{i})$ over all 40 subjects. These correspond to the four covariate patterns of Model 3.

$$
\begin{aligned}
&\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}})\\
&\quad = -2(10)[0.6\ln(0.6/0.4) + \ln(0.4)\\
&\qquad + 0.4\ln(0.4/0.6) + \ln(0.6)\\
&\qquad + 0.3\ln(0.3/0.7) + \ln(0.7)\\
&\qquad + 0.7\ln(0.7/0.3) + \ln(0.3)]\\
&\quad = \mathbf{51.3552}
\end{aligned}
$$

The calculation now shown at the left, where we have substituted each of the four distinct values of $\hat{\text{P}}(\mathbf{X}_\text{i})$ 10 times in the above formula.

$$\boxed{\text{Dev}_{\text{ET}}(\hat{\boldsymbol{\beta}}) = \mathbf{0.0} \neq \text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) = \mathbf{51.3552}}$$

We have thus seen that the events–trial and subject-specific deviance values obtained for Model 3 are numerically quite different.

$\text{Dev}_{\text{ET}}(\hat{\boldsymbol{\beta}}) = 0.0$ because

$$
\begin{aligned}
&-2\ln\hat{L}_{\text{ET saturated}} = -2\ln\hat{L}_{\text{Model 3}}\\
&\quad = -2\ln\hat{L}_{\text{C}}
\end{aligned}
$$

so

$$
\begin{aligned}
\text{Dev}_{\text{ET}}(\hat{\boldsymbol{\beta}}) &= -2\ln\hat{L}_{\text{C}}\\
&\quad - (-2\ln\hat{L}_{\text{ET saturated}})\\
&= -2\ln\hat{L}_{\text{Model 3}}\\
&\quad - (-2\ln\hat{L}_{\text{Model 3}})\\
&= \mathbf{0.0}
\end{aligned}
$$

The reason why $\text{Dev}_{\text{ET}}(\hat{\boldsymbol{\beta}})$ is zero is because the ET formula assumes that the (group-) saturated model is the fully parameterized Model 3 and the current model being considered is also Model 3. So the values of their corresponding log likelihood statistics ($-2\ln\hat{L}$) are equal and their difference is zero.

$\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) \neq 0.0$ because

$$-2\ln\hat{L}_{\text{SS saturated}} = 0.0$$

so

$$
\begin{aligned}
\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) &= -2\ln\hat{L}_{\text{C}}\\
&\quad - (-2\ln\hat{L}_{\text{SS saturated}})\\
&= -2\ln\hat{L}_{\text{Model 3}} - 0\\
&= \mathbf{51.3552}
\end{aligned}
$$

In contrast, the reason why $\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}})$ is different from zero is because the SS formula assumes that the saturated model is the "classical" (SS) saturated model that perfectly predicts each subject's 0 or 1 outcome. As mentioned earlier, $-2\ln\hat{L}$ is always zero for the SS saturated model. Thus, $\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}})$ simplifies to $-2\ln\hat{L}_{\text{C}}$ for the current model (i.e., Model 3), whose value is 51.3552.

Note: $-2\ln\hat{L}_{\text{C,ET}} \neq -2\ln\hat{L}_{\text{C,SS}}$

$-2\ln\hat{L}_{\text{C,ET}} = -2\ln\hat{L}_{\text{C,SS}} - 2K$, where

$$K = \ln\left[\sum_{g=1}^{G}\frac{n_g!}{d_g!(n_g - d_g)!}\right]$$

K does not involve $\boldsymbol{\beta}$, so $\hat{\boldsymbol{\beta}}$ is same for SS or ET

Model 3: $-2\ln\hat{L}_{\text{C,ET}} = 10.8168$

$-2\ln\hat{L}_{\text{C,SS}} = 51.3552, K = 20.2692$

Mathematically, the formula for $-2\ln\hat{L}_{\text{C}}$ *differs with the (ET or SS) format* being used to specify the data. However, these formulae differ by a constant K, as we show on the left and illustrate for Model 3. The formula for K, however, does not involve $\boldsymbol{\beta}$. Thus, the ML estimate $\hat{\boldsymbol{\beta}}$ will be the same for either format. Consequently, some computer packages (e.g., SAS) present the same value (i.e., $-2\ln\hat{L}_{\text{C,SS}}$) regardless of the data layout used.

Deviance not always appropriate for logistic regression GOF

Alternative approach: Hosmer–Lemeshow

We are now ready to discuss why the use of the deviance formula is not always appropriate for assessing GOF for a logistic regression model, and we will describe an alternative approach, using the Hosmer–Lemeshow statistic, which is typically used instead of the deviance.

When $G<<n$, we can assume $\text{Dev}_{\text{ET}}(\hat{\boldsymbol{\beta}})$ is approximately $\chi^2{}_{n-k-1}$ under H_0: good fit

When the number of covariate patterns (G) is considerably smaller than the number of observations (n), the ET deviance formula can be assumed to have an approximate chi-square distribution with $n - k - 1$ degrees of freedom.

EXAMPLE

Previous data: $n = 40$

$G = 2, 4, 4$ for Models $1, 2, 3$, respectively

$\Downarrow$

χ^2 test for GOF is OK

For the data we have illustrated above involving $n = 40$ subjects, $G = 2$ for Model 1, and $G = 4$ for Models 2 and 3. So a chi-square test for GOF is appropriate using the ET deviance formula.

However, **when G ≈ n, we cannot assume** $\text{Dev}_{\text{ET}}(\hat{\boldsymbol{\beta}})$ is approximately $\chi^2{}_{n-p-1}$ under H_0: good fit

(Statistical theory: n_{g} small ≈ 1 as $n \rightarrow \infty$)

However, when G is almost as large as n, in particular, when G equals n, then the deviance cannot be assumed to have a chi-square distribution (Collett, 1991). This follows from large-sample statistical theory, where the primary problem is that in this situation, the number of subjects, n_{g}, for each covariate pattern remains small, e.g., close to 1, as the sample size increases.

X_i continuous, e.g., $X_i = \text{AGE} \Rightarrow G \approx n$

Many situations where predictors are continuous

$\Downarrow$

Cannot use deviance to test for GOF

Note that if at least one of the variables in the model, e.g., AGE, is continuous, then G will tend to be close to n whenever the age range in the sample is reasonably wide. Since logistic regression models typically allow continuous variables, there are many situations in which the chi-square distribution cannot be assumed when using the deviance to test for GOF.

$G = n$:

- Each covariate pattern: 1 subject,
- $n_g \equiv 1$ for all $g, g = 1, \ldots, n$
- $\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) = \text{Dev}_{\text{ET}}(\hat{\boldsymbol{\beta}})$ but not χ^2 under H_0: GOF adequate

When $G = n$, each covariate pattern involves only one subject, and the SS deviance formula is equivalent to the ET deviance formula, which nevertheless cannot be assumed to have a chi-square distribution under H_0.

$$\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) = -2\sum_{i=1}^{n}\left[\hat{\text{P}}(\mathbf{X_i})\ln\left(\frac{\hat{\text{P}}(\mathbf{X_i})}{1-\hat{\text{P}}(\mathbf{X_i})}\right) + \ln\left(1-\hat{\text{P}}(\mathbf{X_i})\right)\right]$$

Provides $\hat{\text{P}}(\mathbf{X}_i)$ **but not** observed Y_i

Moreover, the SS deviance formula, shown again here contains only the *predicted* values $\hat{\text{P}}(\mathbf{X}_i)$ for each subject. Thus, this formula tells nothing about the agreement between *observed* (0,1) outcomes and their corresponding predicted probabilities.

IV. The Hosmer–Lemeshow (HL) Statistic

- Alternative to questionable use of deviance
- Available in most computer packages

To avoid questionable use of the deviance to provide a significance test for assessing GOF, the Hosmer–Lemeshow (HL) statistic has been developed and is available in most computer packages.

HL widely used regardless of whether $G << n$ or $G \approx n$:

- Requires $G > 3$
- Rarely significant when $G < 6$
- Works best when $G \approx n$ (e.g., some Xs are continuous)

The HL statistic is widely used regardless of whether or not the number of covariate patterns (G) is close to the number of observations. Nevertheless, this statistic requires that the model considers at least three covariate patterns, rarely results in significance when G is less than 6, and works best when G is close to n (the latter occurs when some of the predictors are continuous).

- HL $\equiv 0$ for fully parameterized model
- "Saturated" model is fully parameterized in ET format

Moreover, the HL statistic has the property that it will always be zero for a fully parameterized model. In other words, the "saturated" model for the HL statistic is essentially a fully parameterized (group-saturated) model coded in events–trials format.

Steps for computing HL statistic:

1. Compute $\hat{\mathrm{P}}(\mathbf{X}_i)$ for all n subjects
2. Order $\hat{\mathrm{P}}(\mathbf{X}_i)$ from largest to smallest values
3. Divide ordered values into Q percentile groupings (usually $Q = 10 \Rightarrow$ deciles of risk)
4. Form table of observed and expected counts
5. Calculate HL statistic from table
6. Compare computed HL to χ^2 with $Q - 2$ df

The steps involved in computing the HL statistic are summarized at the left. Each step will then be described and illustrated below, following which we will show examples obtained from different models with differing numbers of covariate patterns.

Step 1: Compute $\hat{\mathrm{P}}(\mathbf{X}_i), i = 1, 2, \ldots, n$
$n = 200 \Rightarrow 200$ values for $\hat{\mathrm{P}}(\mathbf{X}_i)$ although some values are identical if $\mathbf{X}_i \equiv \mathbf{X}_j$ for subjects i and j.

At the first step, we *compute the predicted risks* $\hat{\mathrm{P}}(\mathbf{X}_i)$ *for all subjects* in the dataset. If there are, say, $n = 200$ subjects, there will be 200 predicted risks, although some predicted risks will be identical for those subjects with the same covariate pattern.

Step 2: Order values of $\hat{\mathrm{P}}(\mathbf{X}_i)$:
e.g., $n = 200$

Order #	$\hat{\mathrm{P}}(\mathbf{X}_i)$	
1	0.934 (largest)	
2	0.901	tie
3	0.901	
⋮	⋮	
199	0.123	
200	0.045 (smallest)	

At the second step, we *order the predicted risks* from largest to smallest (or smallest to largest).

Again, there may be several ties when doing this if some subjects have the same covariate pattern.

Step 3: Form Q percentile groupings.

Typically, $Q = 10$, i.e., deciles of risk e.g., $n = 200$

Decile	No. of subjects
1	≈ 20
2	≈ 20
$\vdots$	$\vdots$
9	≈ 20
10	≈ 20
Total	200

At the third step, we divide the ordered predicted risks into Q *percentile groupings*. The typical grouping procedure involves $Q = 10$ deciles. Thus, if the sample size is 200, each decile will contain approximately 20 subjects. *Henceforth, we will assume that* $Q = 10$.

Ties ⇒ # of subjects $\neq$ exactly 20 ($= n/Q$) in all deciles

Must keep subjects with identical values of $\hat{\mathrm{P}}(\mathbf{X}_i)$ in the same decile

Note, however, because some subjects may have identical predicted risks (i.e., ties), the number of subjects per decile may vary somewhat to keep subjects with identical predicted risks in the same decile.

Step 4: Form table of observed and expected cases and noncases

Deciles of risk	Obs. cases	Exp. cases	Obs. non cases	Exp. non cases
1	O_{c1}	E_{c1}	O_{nc1}	E_{nc1}
2	O_{c2}	E_{c2}	O_{nc2}	E_{nc2}
3	O_{c3}	E_{c3}	O_{nc3}	E_{nc3}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
10	O_{c10}	E_{c10}	O_{nc10}	E_{nc10}

At the fourth step, we form (typically using a convenient computer program) the table, shown at the left, that contains observed and expected cases and noncases within each decile. In this table, the values O_{cq}, E_{cq}, $O_{\mathrm{nc}q}$, and $E_{\mathrm{nc}q}$, $q = 1, 2, \ldots, 10$ are defined as follows:

O_{cq} = # of observed cases in the qth decile

E_{cq} = # of expected cases in the qth decile

$O_{\mathrm{nc}q}$ = # of observed noncases in the qth decile

$E_{\mathrm{nc}q}$ = # of expected noncases in the qth decile

Observed cases and noncases:

O_{cq} counts # of cases ($Y_i = 1$) in qth decile

$O_{\mathrm{nc}q}$ counts # of noncases ($Y_i = 1$) in qth decile

Note: $O_{\mathrm{nc}q} = n_q - O_{cq}$

The observed cases (O_{cq}) and noncases ($O_{\mathrm{nc}q}$) in each decile are obtained by simply counting the numbers of subjects in that decile who are cases (i.e., $Y_i = 1$) and noncases (i.e., $Y_i = 0$), respectively. Note that once we count O_{cq}, we can obtain $O_{\mathrm{nc}q}$ by subtraction from n_q, the total number of subjects in the qth decile.

Expected cases and noncases:

$E_{cq} = \sum_{i=1}^{n_q} \hat{\mathrm{P}}(\mathbf{X}_{iq})$ and $E_{\mathrm{nc}q} = n_q - E_{cq}$, where

$\mathbf{X}_{iq}$ = covariate values for ith subj in qth decile

The expected cases (E_{cq}) in each decile are obtained by summing the predicted risks $\hat{\mathrm{P}}(\mathbf{X}_i)$ for all subjects in that decile. The expected number of noncases ($E_{\mathrm{nc}q}$) are obtained by subtraction from n_{q}.

EXAMPLE

e.g., $q = 3$, $n_3 = 4$, $\hat{\mathrm{P}}(\mathbf{X}_i)$: 0.30, 0.35, 0.40, 0.45

$$E_{c3} = \sum_{i=1}^{n_3} \hat{\mathrm{P}}(\mathbf{X}_{i3}) = 0.30 + 0.35 + 0.40 + 0.45 = \mathbf{1.50}$$

and $E_{nc3} = \mathrm{n}_3 - \mathrm{E}_{c3} = 4 - 1.50 = \mathbf{2.50}$

For example, if the third decile contains four subjects with predicted risks of 0.30, 0.35, 0.40, and 0.45, then the expected number of cases (E_{c3}) would be their sum $0.30 + 0.35 + 0.40 + 0.45 = 1.50$ (regardless of whether or not a subject is an observed case). The expected noncases in the same decile (E_{nc3}) would be $4 - 1.50 = 2.50$.

Step 5:

$$\boxed{\mathrm{HL} = \sum_{q=1}^{Q} \frac{(\mathbf{O}_{cq} - \mathbf{E}_{cq})^2}{\mathbf{E}_{cq}} + \sum_{q=1}^{Q} \frac{(\mathbf{O}_{ncq} - \mathbf{E}_{ncq})^2}{\mathbf{E}_{ncq}}}$$

$Q = 10 \Rightarrow 20$ values in summation

In step 5, the HL statistic is calculated using the formula at the left. This formula involves summing Q values of the general form $(O_q - E_q)^2/E_q$ for cases and another Q values for noncases. When $Q = 10$, the HL statistic therefore involves 20 values in the summation.

Step 6:

HL ~ approx χ^2_{Q-2} under H_0: Good fit

$Q = 10 \Rightarrow \mathrm{df} = Q - 2 = 8$

i.e., not enough evidence to indicate lack of fit

In step 6, the HL statistic is tested for significance by comparing the computed HL value to a percentage point of χ^2 with $Q - 2$ degrees of freedom. When $Q = 10$, therefore, the HL statistic is approximately χ^2 with 8 df.

V. Examples of the HL Statistic

EXAMPLE

Evans County Data ($n = 609$) (see previous chapters and Computer Appendix)

Model EC1 (no interaction):

$$\text{logit P}(\mathbf{X}) = \alpha + \beta\text{CAT} + \gamma_1\text{AGEG} + \gamma_2\text{ECG}$$

Model EC2 (fully parameterized):

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha &+ \beta\text{CAT} + \gamma_1\text{AGEG}\\ &+ \gamma_2\text{ECG}\\ &+ \gamma_1\text{AGEG} \times \text{ECG}\\ &+ \delta_1\text{CAT} \times \text{AGE}\\ &+ \delta_2\text{CAT} \times \text{ECG}\\ &+ \delta_3\text{CAT} \times \text{AGE} \times \text{ECG}\end{aligned}$$

We now illustrate the use of the HL statistic with the Evans County data ($n = 609$). This dataset has been considered in previous chapters, and is described in detail in the Computer Appendix. SASs Logistic procedure was used for the computations.

In our first illustration, we fit the two models shown at the left. The outcome variable is CHD status (1 = case, 0 = noncase), and there are three basic (i.e., main effect) binary predictors, CAT (1 = high, 0 = low), AGEG (1 = age $\leq$ 55, 0 = age > 55), and ECG (1 = abnormal, 0 = normal). Recall that the Evan County dataset is described in the Computer Appendix.

EXAMPLE (continued)

Datalines in events trials format (n = 609)

Group (g)	d_g	n_g	CAT	AGEG	ECG
1	17	274	0	0	0
2	15	122	0	1	0
3	7	59	0	0	1
4	5	32	0	1	1
5	1	8	1	0	0
6	9	39	1	1	0
7	3	17	1	0	1
8	14	58	1	1	1

The data layout used to fit both models in events–trials format is now shown at the left. Model EC1 is a no-interaction model involving only the main effects CAT, AGEG, and ECG as predictors. Model EC2 is a fully parameterized model since there are eight model parameters as well as eight covariate patterns, i.e., $p + 1 = 8 = G$.

```
proc logistic data = evans2;
model cases/total = cat ageg ecg/
scale = none aggregate = (cat
ageg ecg) lackfit;
output out = pred p = phat
predprob = (individual); run;
proc print data = pred; run;
```

Here, we provide the computer code using SASs PROC LOGISTIC used to fit Model EC1 and provide HL, deviance, and $-2\ln\hat{L}_C$ statistics, as well as predicted risk values (i.e., "phat" in the code at the left) for each covariate pattern.

EXAMPLE

Edited Output (Model EC1):
(Variables – CAT, AGE, ECG)

```
Partition for the Hosmer and Lemeshow Test

Pct                  Event        Nonevent
Grp (q)   Total  Obs     Exp    Obs       Exp
 1          274   17   18.66    257    255.34
 2           59    7    5.60     52     53.40
 3          122   15   14.53    107    107.47
 4           57    9    8.94     48     48.06
 5           39    9    7.85     30     31.15
 6           58   14   15.41     44     42.59

 Hosmer and Lemeshow  Goodness-of-Fit Test
              Chi-Square       DF     Pr > ChiSq
                  0.9474        4         0.9177

 Deviance         0.9544        4         0.9166
  Pearson         0.9793        4         0.9129

 −2 Log L        418.181

                          Probabilities
                          Pred       Obs
        Group (g)         phat        p
            1          0.06810    0.06204
            2         ┌0.11913    0.12295
            3         │0.09497    0.11864
            4         │0.16264┐   0.15625
            5         └0.11984│   0.12500
            6          0.20128│   0.23077
            7          0.16355┘   0.17647
            8          0.26574    0.24138
```

No evidence that Model EC1 has lack of fit

Edited output for Model EC1 is shown here.

The table of observed and expected cases and noncases has divided the data *into Q = 6 percentile groups* rather than 10 deciles. The number of covariate patterns is $G = 8$, so the number of percentile groups allowable is less than 10. Also, from the *phat* values provided (below left), two pairs {0.11913, 0.11984} and {0.16264, 0.16355} are essentially identical and should not be separated into different groups.

Since $Q = 6$, the df for the HL test is $Q - 2 = 4$. The HL test statistic (0.9474) is not significant. Thus, there is not enough evidence to indicate that Model EC1 has lack of fit.

The output here also gives the *Deviance* (0.9544) and *Pearson* (0.9793) chi-square statistics as well as the log likelihood statistic (418.181) for Model EC1.

EXAMPLE (continued)

Since $G = 8 << n = 609$,
Pearson statistic and Dev statistic approx χ^2 under H_0

The *Pearson statistic* is another GOF statistic that is similar to the Deviance in that it is not recommended when the number of covariate patterns is close to the sample size. However, since the number of covariate patterns ($G = 8$) for Model EC1 is much less than the sample size ($n = 609$), both statistics can be assumed to be approximately chi square under H_0. Notice also that the Pearson and Deviance values are very close to the HL value (0.9474).

$$-2\ln \hat{L}_{\text{C,SS}} = 418.181$$
$$= \text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) \text{ for Model EC1}$$

The $-2 \log L$ value (418.181) in the output is the SS statistic $-2\ln \hat{L}_{\text{C,SS}}$, where C is Model EC1. This statistic is equivalent to $\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}})$ for Model EC1, since $\hat{L}_{\text{max,SS}}$ is always one.

$$\text{Dev}_{\text{ET}}(\hat{\beta}) = 0.9544 = \underbrace{-2 \ln \hat{L}_{\text{EC1,SS}}}_{418{,}181} - \underbrace{(-2 \ln \hat{L}_{\text{EC2,SS}})}_{417{,}226}$$

The Deviance in the output (0.9544) is computed using the $\text{Dev}_{\text{ET}}(\hat{\boldsymbol{\beta}})$ formula based on the ET data layout. This formula is also equivalent to the difference between SS log-likelihood statistics for Model EC1 (418.181) and the (fully parameterized) Model EC2 (417.226, in output below).

Table of Probabilities (**phat** vs. **p**)
$\Downarrow$
No perfect group prediction

e.g., group 3: phat = 0.09497
p = 0.11864

Also, from the table of probabilities (above left), the observed and predicted probabilities are different, so Model EC1 does not provide perfect group (i.e., covariate pattern) prediction.

EXAMPLE

```
Edited Output (Model EC2):
(Variables – CAT, AGE, ECG, AGE × ECG, CAT × AGE,
AGE × ECG, and CAT × AGE × ECG)
Partition for the Hosmer and Lemeshow Test

Pct              Event           Nonevent
Grp (q)  Total  Obs     Exp     Obs      Exp
 1        274   17    17.00     257   257.00
 2         59    7     7.00      52    52.00
 3        122   15    15.00     107   107.00
 4         57    9     9.00      48    48.00
 5         39    9     9.00      30    30.00
 6         58   14    14.00      44    44.00

   Hosmer and Lemeshow Goodness-of-Fit Test
           Chi-Square      DF     Pr > ChiSq
             0.0000         4       1.0000

  Deviance     0.0000        4
  Pearson      0.0000        4

 −2 Log L    417.226

                      Probabilities
                     Pred        Obs
     Group           phat         p
       1           0.06205     0.06205
       2           0.12295     0.12295
       3           0.11864     0.11864
       4           0.15625     0.15625
       5           .012500     0.12500
       6           0.23079     0.23079
       7           0.17647     0.17647
       8           0.24138     0.24138
```

We now show edited ouput for the fully parameterized Model EC2.

As with Model EC1, Model EC2 has eight covariate patterns, and only $Q = 6$ percentile groups are obtained in the table of observed and expected cases and noncases. However, since Model EC2 is fully parameterized ($k + 1 = 8 = G$), corresponding observed and expected cases and noncases are identical throughout the table.

Consequently, the HL test statistic is zero, as are both the Deviance and Pearson statistics.

The log likelihood statistic of 417.226 is equivalent to the SS deviance (i.e., $\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}})$) for Model EC2. Since this deviance value is different from 0, we know that Model EC2 is not the (SS) saturated model that perfectly predicts the 0 or 1 outcome for each of the 609 subjects in the dataset.

Pred = obs for each g

(phat = p)

$$\Downarrow$$

$\hat{d}_g = d_g$ cases

Yet, as the table of probabilities indicates, Model EC2 perfectly predicts the observed probabilities obtained for each covariate pattern. Equivalently, then, this model perfectly predicts the observed number of cases (d_g) corresponding to each covariate pattern.

EXAMPLE

Fully parameterized model (w/o conts. Xs):

- Above results support its use as gold standard for assessing GOF
- HL statistic always 0 when perfect group prediction

When none of the predictors are continuous, as with these data, these results support the use of a fully parameterized model defined from all covariate patterns as the gold standard model for assessing GOF. In particular, the HL statistic reflects this framework, since the value of HL will always be zero whenever there is perfect group, rather than subject-specific, prediction.

Second illustration (Evans County data):
Model EC3 (no interaction):

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha &+ \beta\text{CAT} + \gamma_1\text{AGE}\\ &+ \gamma_2\text{ECG} + \gamma_3\text{SMK}\\ &+ \gamma_4\text{CHL} + \gamma_5\text{HPT}\end{aligned}$$

Model EC4:

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha &+ \beta\text{CAT} + \gamma_1\text{AGE}\\ &+ \gamma_2\text{ECG} + \gamma_3\text{SMK}\\ &+ \gamma_4\text{CHL} + \gamma_5\text{HPT}\\ &+ \delta_1\text{CAT} \times \text{CHL}\\ &+ \delta_2\text{CAT} \times \text{HPT}\end{aligned}$$

We now provide a second illustration of GOF assessment using the Evans County data (see Computer Appendix) with models that involve continuous variables. In particular, we consider two previously considered models (see Chap. 7) shown on the left that involve the predictors CAT, AGE, ECG, SMK, CHL, and HPT. Here, AGE and CHL are continuous, whereas CAT, ECG, SMK, and HPT are binary variables.

EXAMPLE

AGE and CHL continuous

$$\Downarrow$$

Few subjects with identical values for both AGE and CHL

$$\Downarrow$$

$G \approx n(= 609)$

Since both models EC3 and EC4 contain continuous variables AGE and CHL, there are not likely to be many of the 609 subjects with identically the same values for these two variables. Consequently, the number of covariate patterns (G) for each model should be close to the sample size of 609.

SASs Proc Logistic automatically outputs "Number of Unique Profiles" (G)

Here, $G = 599$

In fact, SASs Logistic procedure automatically outputs this number (identified in the output as the "number of unique profiles"), which turns out to be 599 (see output, below left), i.e., $G = 599$.

EXAMPLE

$G = 599 \approx n\ (= 609)$:

- Deviance nor Pearson not approximately $\sim \chi^2$
- HL statistic approximately $\sim \chi^2$ (can use HL for GOF test)

Since $G \approx n$ in both models, we therefore cannot assume that the Deviance or Pearson statistics are approximately chi square. However, we can use the Hosmer–Lemeshow (HL) statistic to carry out a test for GOF.

```
Edited Output (Model EC3):
(Variables–CAT, AGE, ECG, SMK, CHL, and HPT)

Partition for the Hosmer and Lemeshow Test

Pct                  Event          Nonevent
Grp (q)  Total  Obs    Exp    Obs       Exp
 1        61     0     1.72   61      59.28
 2        61     2     2.65   59      58.35
 3        61     5     3.46   56      57.54
 4        61     6     4.22   55      56.78
 5        61     7     5.21   54      55.79
 6        61     6     6.10   55      54.90
 7        61     5     7.50   56      53.50
 8        61     9     9.19   52      51.81
 9        61    12    11.86   49      49.14
10        60    19    19.08   41      40.92


   Hosmer and Lemeshow Goodness-of-Fit Test
               Chi-Square     DF    Pr > ChiSq
                 5.1028        8      0.7465

 Deviance       400.3938     592      1.0000
 Pearson        589.6446     592      0.5196

Number of unique profiles: 599

 −2 Log L       400.394
```

No evidence Model EC3 has lack of fit

On the left, we now show edited output for Model EC3, which provides the HL information as well as Deviance, Pearson, and −2 log L statistics.

From the output for Model EC3, we see that the predicted risks have been divided into ($Q = 10$) deciles, with about 61 subjects in each decile. Also, the observed and expected cases are somewhat different within each decile, and the observed and expected noncases are somewhat different.

The HL statistic of 5.1028 has $Q - 2 = 8$ degrees of freedom and is nonsignificant ($P = 0.7465$). Thus, there is no evidence from this test that the no-interaction Model EC3 has lack of fit.

Both the Deviance (400.3938) and Pearson (589.6446) statistics are very different from each other as well as from the HL statistic (5.1028). This is not surprising since the Deviance and Pearson statistics are not appropriate for GOF testing here.

$$-2 \text{ Log } L = \text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) = -2\ln(\hat{L}_{\text{SS,EC3}}/\hat{L}_{\text{SS,max}})$$

$$\text{since Log } \hat{L}_{\text{SS,max}} \equiv 0$$

Note that since the log likelihood for the (SS) saturated model is always zero, the log likelihood (−2 Log L) value of 400.394 is identical to the Deviance value (i.e., $\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}})$) for Model EC3. We will use this value of −2 Log L in an LR statistic (below) that compares the no-interaction Model E3 to the interaction Model EC4.

EXAMPLE

Edited Output (Model EC4):
(Variables – CAT, AGE, ECG, SMK, CHL, HPT
CAT × CHL, and CAT × HPT)

```
Partition for the Hosmer and Lemeshow Test
Pct                 Event          Nonevent
Grp (q)   Total  Obs     Exp     Obs       Exp
 1          61     2     0.94    59       60.06
 2          61     1     1.96    60       59.04
 3          61     2     2.68    59       58.32
 4          61     5     3.37    56       57.63
 5          61     4     4.07    57       56.93
 6          61     2     4.78    59       56.22
 7          61     4     5.77    57       55.23
 8          61    12     7.66    49       53.34
 9          61    10    11.05    51       49.95
10          60    29    28.71    31       31.29


  Hosmer and Lemeshow Goodness-of-Fit Test
             Chi-Square        DF      Pr > ChiSq
               7.9914           8        0.4343

  Deviance     347.2295        590       1.0000
  Pearson      799.0580        590       <.0001

  −2 Log L     347.230
```

No evidence
Model EC4
has lack of fit

Edited output for the interaction model EC4 is now shown here at the left. This model contains CAT, AGE, ECG, SMK, CHL, and HPT, and the product terms CAT × CHL and CAT × HPT in addition to the six main effects.

From this output, as with Model EC3, we see that the predicted risks have been divided into ($Q = 10$) deciles, with about 61 subjects in each decile. Also, the observed and expected cases are somewhat different within each decile, and the observed and expected noncases are somewhat different.

The HL statistic of 7.9914 has $Q - 2 = 8$ degrees of freedom and is nonsignificant ($P = 0.4343$). Thus, there is not sufficient evidence from this test to conclude that interaction Model EC4 has lack of fit (as concluded for Model EC3).

Deviance statistic (347.2295)
very different from
Pearson statistic (799.0580)

As with Model EC3, both the Deviance (347.2295) and Pearson (799.0580) statistics for Model EC4 are very different from each other as well as from the HL statistic (7.9914).

Both Models EC3 and EC4
do not have lack of fit:

Use LR test to compare Models EC3 vs. EC4

Although we can conclude from the HL test that both Models EC3 and EC4 do not indicate lack of fit, we can decide between these two models by performing an LR test that compares corresponding log likelihood statistics for the two models.

H_0: $\delta_1 = \delta_2 = 0$ in Model EC4

$$\text{LR} = -2\ln\hat{L}_{\text{EC3}} - (-2\ln\hat{L}_{\text{EC4}}) \sim \chi^2_{2\text{ df}} \text{ under } H_0$$

The null hypothesis is that the coefficients of the two product terms in Model EC4 are both zero. As previously seen in Chap. 7, the test statistic is approximately chi square with 2 degrees of freedom under the null hypothesis.

	Model EC3	Model EC4
-2 Log L	400.394	347.230

$\text{LR} = 400.394 - 347.230 = 53.165$
($P < 0.001$):
Model EC4 preferred over Model EC3

The resulting LR value is 53.164, which is highly significant ($P < 0.0001$).

Consequently, the interaction Model EC4 is preferred to the no-interaction Model EC3.

VI. SUMMARY

✓ Chapter 9: Assessing Goodness of Fit for Logistic Regression

This presentation is now complete. We have described how to assess the extent to which a binary logistic model of interest predicts the observed outcomes in one's dataset.

Saturated model:

- Contains as many parameters ($p + 1$) as the number of subjects (n) in the dataset
- Provides perfect prediction of the observed (0, 1) outcomes on each subject

Fully parameterized model:

- Contains the maximum number of covariates that can be defined from the basic predictors ($\mathbf{X}$) being considered for the model
- Provides perfect prediction of the observed proportion of cases within subgroups defined by distinct covariate patterns of $\mathbf{X}$

We have identified two alternative models, a *saturated model* and a *fully parameterized model*, that can be used as possible gold standard referent points for evaluating the fit of a given model.

Subject-specific (SS) format:

- Datalines listed by subjects
- Used for GOF measure of model fit for (0, 1) outcomes

Events–trials (ET) format:

- Datalines listed by subgroups based on (G) covariate patterns
- Used for GOF measure of model fit for subgroup proportions

We have also distinguished between two alternative data layouts that can be used – *subject specific* (SS) vs. *events–trials* (ET) formats.

Deviance:

- Likelihood ratio (LR) statistic for comparing one's current model to the saturated model
- Not recommended when $G \approx n$

A widely used GOF measure for many mathematical models is called the *deviance*. However, the deviance is not recommended for a binary logistic regression model in which the number of covariate patterns (G) is close to the number of subjects (n).

Hosmer–Lemeshow (HL) statistic:

- GOF statistic appropriate when $G \approx n$
- Computed using O and E cases and noncases in percentile subgroups

In the latter situation, a popular alternative is the Hosmer–Lemeshow (*HL*) statistic, which is computed from a table of observed and expected cases and noncases categorized by percentile subgroups, e.g., deciles of predicted probabilities.

We suggest that you review the material covered here by reading the detailed outline that follows. Then do the practice exercises and test.

Chapter 10: Assessing Discriminatory Performance of a Binary Logistic Model

In the next chapter (Chap. 10), we describe methods for assessing the discriminatory performance is of a binary logistic model using misclassification tables and ROC curves.

VII. Appendix: Derivation of the Subject-Specific (SS) Deviance Formula

$$\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) = -2\sum_{k=1}^{n}\left[\hat{\text{P}}(\mathbf{X_i})\ln\left(\frac{\hat{\text{P}}(\mathbf{X_k})}{1-\hat{\text{P}}(\mathbf{X_k})}\right) + \ln\left(1-\hat{\text{P}}(\mathbf{X_k})\right)\right]$$

Proof. We first write the Deviance formula in a convenient form as follows:

$$\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) = -2\ln\left(\frac{\hat{L}_{\text{C}}}{\hat{L}_{\text{MAX}}}\right)$$

$$= -2\ln\hat{L}_{\text{C}} \text{ since } \ln\hat{L}_{\text{MAX}} \equiv 0$$

$$-2\ln,\hat{L}_{\text{C}} \overset{\text{definition}}{=} -2\sum_{k=1}^{n}\Big[Y_k\ \ln\hat{\text{P}}(\mathbf{X}_k). + (1-Y_k)\ln(1-\hat{\text{P}}(\mathbf{X}_k)\Big]$$

$$\overset{\text{algebra}}{=} -2\sum_{k=1}^{n}\left[Y_k\ \ln\left(\frac{\hat{\text{P}}(\mathbf{X}_k)}{1-\hat{\text{P}}(\mathbf{X}_k)}\right) + \ln(1-\hat{\text{P}}(\mathbf{X}_k)\right].$$

We now write the log of the logistic likelihood function in a convenient form and take its derivative:

$$L(\boldsymbol{\beta}) \overset{\text{definition}}{=} \prod_k \hat{\text{P}}(\mathbf{X}_k)^{Y_k}(1-\hat{\text{P}}(\mathbf{X}_k))^{1-Y_k}$$

so

$$\ln\ L(\boldsymbol{\beta}) = \sum_k Y_k\ln\hat{\text{P}}(\mathbf{X}_k) + (1-Y_k)\ln(1-\hat{\text{P}}(\mathbf{X}_k)).$$

Taking derivatives, we obtain

$$\frac{\partial \ln\ L(\boldsymbol{\beta})}{\partial \beta_{\mathrm{j}}} = \sum_k \left\{ \frac{Y_k}{\hat{\mathrm{P}}(\mathbf{X}_k)} - \frac{1 - Y_k}{1 - \hat{\mathrm{P}}(\mathbf{X}_k)} \right\} \hat{\mathrm{P}}(\mathbf{X}_k) \times (1 - \hat{\mathrm{P}}(\mathbf{X}_k))X_{\mathrm{j}k},$$

which can be rewritten as

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_{\mathrm{j}}} = \sum_k \left\{ Y_k(1 - \hat{\mathrm{P}}(\mathbf{X}_k)) - (1 - Y_k)\hat{\mathrm{P}}(\mathbf{X}_k) \right\} X_{\mathrm{j}k}$$

and further simplied to

$$\frac{\partial \ln\ L(\boldsymbol{\beta})}{\partial \beta_{\mathrm{j}}} = \sum_k (Y_k - \hat{\mathrm{P}}(\mathbf{X}_k))X_{\mathrm{jk}}.$$

We can then write

$$\begin{aligned} \sum_{\mathrm{j}} \beta_{\mathrm{j}} \frac{\partial\ \ln\ L(\boldsymbol{\beta})}{\partial \beta_{\mathrm{j}}} &= \sum_k \left(Y_k - \hat{\mathrm{P}}(\mathbf{X}_k) \right) \sum_{\mathrm{j}} \beta_{\mathrm{j}} X_{\mathrm{j}k} \\ &= \sum_k \left(Y_k - \hat{\mathrm{P}}(\mathbf{X}_k) \right) \ln \left(\frac{\hat{\mathrm{P}}(\mathbf{X}_k)}{1 - \hat{\mathrm{P}}(\mathbf{X}_k)} \right) \\ &= \sum_k \left(Y_k - \hat{\mathrm{P}}(\mathbf{X}_k) \right) \mathrm{logit}\ \hat{\mathbf{P}}(\mathbf{X}_{\mathrm{k}}). \end{aligned}$$

Since $\frac{\partial\ \ln\ L(\hat{\boldsymbol{\beta}})}{\partial \beta_{\mathrm{j}}} = 0$ for the ML estimate $\hat{\boldsymbol{\beta}}$, we can write $\sum_k (Y_k - \hat{\mathrm{P}}(\mathbf{X}_k))\ \mathrm{logit}\ \hat{\mathrm{P}}(\mathbf{X}_k) = 0$.

It then follows that $\sum_k Y_k\, \mathrm{logit}(\hat{\mathrm{P}}(\mathbf{X}_k)) = \sum_k \hat{\mathrm{P}}(\mathbf{X}_k)\, \mathrm{logit}(\hat{\mathrm{P}}(\mathbf{X}_k)^k)$.

We then replace $\sum_k Y_k\, \mathrm{logit}(\hat{\mathrm{P}}(\mathbf{X}_k))$ by $\sum_k \hat{\mathrm{P}}(\mathbf{X}_k)\, \mathrm{logit}(\hat{\mathrm{P}}(\mathbf{X}_k))$ in the above simplified formula for the deviance to obtain

$$\mathrm{Dev}_{\mathrm{SS}}(\hat{\boldsymbol{\beta}}) = -2 \sum_{k=1}^{n} \left[\hat{\mathrm{P}}(\mathbf{X}_k)\, \mathrm{logit}(\hat{\mathrm{P}}(\mathbf{X}_k)) + \ln(1 - \hat{\mathrm{P}}(\mathbf{X}_k)) \right].$$

Detailed Outline

I. Overview (pages 304–305)

A. Focus: Goodness of fit (GOF) – assessing the extent to which a logistic model estimated from a dataset predicts the observed outcomes in the dataset.

B. Considers how well a given model, considered by itself, fits the data.

C. Provides a summary measure over all subjects that compares the observed outcome (Y_i) for subject i to the predicted outcome ($\hat{Y}_i$) for this subject obtained from the fitted model.

D. Widely used measure is the *deviance*; however, for binary logistic regression, use of deviance is problematic. Alternative measure: *Hosmer–Lemeshow* (HL) statistic.

II. Saturated vs. Fully Parameterized Models (pages 305–312)

A. Saturated model

i. Provides *perfect prediction of the (0, 1) outcome* for each subject in the dataset

ii. Contains as many parameters as the number of "subjects" in the dataset

iii. Uses data layout in *subjects-specific (SS) format*

iv. Classical model used as gold standard for assessing GOF

B. Fully parameterized model

i. Contains the maximum number of covariates that can be defined from the basic predictors (**X**) being considered for the model.

ii. The number of parameters ($k + 1$) equals the number (G) of distinct *covariate patterns* (or *subgroups*) that can be defined from the basic predictors.

iii. Uses data layout in *events–trials (ET) format*.

iv. Provides *perfect prediction* of the observed proportion of cases within *subgroups* defined by distinct covariate patterns of **X**.

v. An alternative gold standard model for determining GOF.

C. Example: $n = 40$, 2 basic predictors: E (0, 1), V(0, 1)

i. Fully parameterized model ($G = 4$ covariate patterns, $k = 3$ variables):

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \gamma V + \delta EV$$

ii. Saturated model ($n = 40$ parameters)

$$\text{logit P}(\mathbf{X}) = \omega_1 Z_1 + \omega_2 Z_2 + \omega_3 Z_3 + \cdots + \omega_{40} Z_{40}$$

$$Z_i = \left\{ \begin{array}{l} 1 \text{ if subject } i; \quad i = 1, 2 \ldots, 40 \\ 0 \text{ otherwise} \end{array} \right] .$$

III. The Deviance Statistic (pages 312–317)

A. Formula: $\text{Dev}(\hat{\boldsymbol{\beta}}) = -2\ln(\hat{L}_c/\hat{L}_{\max})$, where
$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$
$\hat{L}_c$ = ML for current model
$\hat{L}_{\max}$ = ML for saturated model

B. Contrasts the likelihood of the current model with the likelihood of the model that perfectly predicts the observed outcomes

C. The closer $\hat{L}_c$ and $\hat{L}_{\max}$ are to one another, the better the fit (and the smaller the deviance

D. Common to test for GOF by comparing deviance with χ^2_{n-p-1} value, but *questionably legitimate*

E. There are two alternative formulae for the deviance:

i. $\text{Dev}_{\text{ET}}(\hat{\boldsymbol{\beta}})$

$$= -2\sum_{g=1}^{G}\left[d_g \ln\left(\frac{d_g}{\hat{d}_g}\right) + (n_g - d_g)\ln\left(\frac{n_g - d_g}{n_g - \hat{d}_g}\right)\right]$$

uses events–trials format, where

$\hat{d}_g = n_g \hat{\text{P}}(\mathbf{X}_g)$ = # of expected cases,

d_g = # of observed cases,
G = # of covariate patterns

ii. $\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}})$

$$= -2\sum_{i=1}^{n}\left[Y_i \ln\left(\frac{Y_i}{\hat{Y}_i}\right) + (1 - Y_i)\ln\left(\frac{1-Y_i}{1-\hat{Y}_i}\right)\right]$$

uses subject-specific format, where

Y_i = observed (0, 1) response for subject i
and $\hat{Y}_i$ = predicted probability for subject $i = \hat{P}(\mathbf{X}_i)$

iii. $\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}}) \neq \text{Dev}_{\text{ET}}(\hat{\boldsymbol{\beta}})$ unless $G = n$

iv. Fully parameterized model:

$\text{Dev}_{\text{ET}}(\hat{\boldsymbol{\beta}}) = 0$ always but

$\text{Dev}_{\text{SS}}(\hat{\boldsymbol{\beta}})$ is never 0

F. Using $\mathrm{Dev}_{\mathrm{ET}}(\hat{\boldsymbol{\beta}})$ to test for GOF:

  i. When $G << n$, can assume $\mathrm{Dev}_{\mathrm{ET}}(\hat{\boldsymbol{\beta}})$ is approximately $\chi^2{}_{n-\mathrm{p}-1}$ under H_0: good fit

  ii. However, when $G \approx n$, *cannot* assume $\mathrm{Dev}_{\mathrm{ET}}(\hat{\boldsymbol{\beta}})$ is approximately $\chi^2{}_{n-\mathrm{p}-1}$ under H_0: good fit

  iii. X_i continuous, e.g., $X_i = \mathrm{AGE} \Rightarrow G \approx n$, so cannot test for GOF

G. Why $\mathrm{Dev}_{\mathrm{SS}}(\hat{\boldsymbol{\beta}})$ cannot be used to test for GOF:

  i. Alternative formula for $\mathrm{Dev}_{\mathrm{SS}}(\hat{\beta})$: $\mathrm{Dev}_{\mathrm{SS}}(\hat{\boldsymbol{\beta}})$
$$= -2\sum_{i=1}^{n}\left[\hat{\mathrm{P}}(\mathbf{X_i})\ln\left(\frac{\hat{\mathrm{P}}(\mathbf{X_i})}{1-\hat{\mathrm{P}}(\mathbf{X_i})}\right)+\ln\left(1-\hat{\mathrm{P}}(\mathbf{X_i})\right)\right]$$

  ii. The above formula contains only the *predicted* values $\hat{\mathrm{P}}(\mathbf{X}_i)$ for each subject; tells nothing about the agreement between *observed* (0, 1) outcomes and their corresponding predicted probabilities

IV. The HL Statistic (pages 318–320)

A. Used to provide a significance test for assessing GOF:

  i. Avoids questionable use of the deviance when $G \approx n$

  ii. Available in most computer procedures for logistic regression

  iii. Requires that the model considers at least three covariate patterns, rarely results in significance when G is less than 6, and works best when G is close to n, e.g., with continuous predictors

B. Steps for computation:

  1. Compute $\hat{\mathrm{P}}(\mathbf{X}_i)$ for all n subjects
  2. Order $\hat{\mathrm{P}}(\mathbf{X}_i)$ from largest to smallest values
  3. Divide ordered values into Q percentile groupings (usually $Q = 10$, i.e., deciles)
  4. Form table of observed and expected counts
  5. Calculate HL statistic from table
  6. Compare computed HL to χ^2 with $Q - 2$ df

C. Table of observed and expect counts (Step 4)

Deciles of risk	Obs. cases	Exp. cases	Obs. non cases	Exp. non cases
1	O_{c1}	E_{c1}	O_{nc1}	E_{nc1}
2	O_{c2}	E_{c2}	O_{nc2}	E_{nc2}
3	O_{c3}	E_{c3}	O_{nc3}	E_{nc3}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
10	O_{c10}	E_{c10}	O_{nc10}	E_{nc10}

D. HL Statistic formula (Step 5):

$$\text{HL} = \sum_{q=1}^{Q} \frac{(\mathbf{O}_{cq} - \mathbf{E}_{cq})^2}{\mathbf{E}_{cq}} + \sum_{q=1}^{Q} \frac{(\mathbf{O}_{ncq} - \mathbf{E}_{ncq})^2}{\mathbf{E}_{ncq}}$$

V. Examples of the HL Statistic (pages 320–325)

A. Two examples, each using the Evans County data ($n = 609$).

B. Example 1 uses two models involving three binary predictors with data layout in events–trials format ($G = 8$).

i. The models

Model EC1 (no interaction): logit $P(\mathbf{X}) = \alpha + \beta\text{CAT} + \gamma_1\text{AGEG} + \gamma_2\text{ECG}$

Model EC2: logit $P(\mathbf{X}) = \alpha + \beta\text{CAT} + \gamma_1\text{AGEG} + \gamma_2\text{ECG} + \gamma_3\text{AGEG} \times \text{ECG} + \delta_1\text{CAT} \times \text{AGE} + \delta_2\text{CAT} \times \text{ECG} + \delta_3\text{CAT} \times \text{AGE} \times \text{ECG}$

ii. Model EC2 is fully parameterized, which, as expected, perfectly predicts the observed number of cases (d_g) corresponding to each covariate pattern.

iii. The HL test statistic for Model EC2 is zero.

C. Example 2 uses two models that involve continuous variables.

i. The models:

Model EC3 (no interaction): logit $P(\mathbf{X}) = \alpha + \beta\text{CAT} + \gamma_1\text{AGE} + \gamma_2\text{ECG} + \gamma_3\text{SMK} + \gamma_4\text{CHL} + \gamma_5\text{HPT}$

Model EC4: logit $P(\mathbf{X}) = \alpha + \beta\text{CAT} + \gamma_1\text{AGE} + \gamma_2\text{ECG} + \gamma_3\text{SMK} + \gamma_4\text{CHL} + \gamma_5\text{HPT} + \delta_1\text{CAT} \times \text{CHL} + \delta_2\text{CAT} \times \text{HPT}$.

ii. The number of covariate patterns (G) for each model is 599, which is quite close to the sample size (n) of 609.

iii. The HL statistics computed for both models are not significant, i.e., there is no evidence from this test that either model has lack of fit.

iv. Can decide between Models EC3 and EC4 by performing an LR test that corresponding log likelihood statistics or deviances for the two models.

VI. Summary (page 326)

VII. Appendix: Derivation of SS Deviance Formula (pages 327–328)

Practice Exercises

The following questions and computer information consider the Evans Country dataset on 609 white males that has been previously discussed and illustrated in earlier chapters of this text. Recall that the outcome variable is CHD status (1 = case, 0 = noncase), the exposure variable of interest is CAT status (1 = high CAT, 0 = low CAT). In this example, we consider only two categorical control variables AGEG (1 = age > 55, 0 = age ≤ 55) and ECG (1 = abnormal, 0 = normal). The dataset involving the above variables is given as follows:

Cases	Total	CAT	AGE	ECG
17	274	0	0	0
15	122	0	1	0
7	59	0	0	1
5	32	0	1	1
1	8	1	0	0
9	39	1	1	0
3	17	1	0	1
14	58	1	1	1

The SAS output provided below was obtained for the following logistic model:

$$\text{Logit P}(\mathbf{X}) = \alpha + \beta_1\text{CAT} + \gamma_1\text{AGE} + \gamma_2\text{ECG}$$

```
     Deviance and Pearson Goodness-of-Fit Statistics

Criterion     Value     DF    Value/DF     Pr > ChiSq
Deviance      0.9544    4      0.2386        0.9166
Pearson       0.9793    4      0.2448        0.9129

             Number of unique profiles: 8

                  Model Fit Statistics

Criterion    Intercept Only   Intercept and Covariates
-2 Log L        438.558                418.181

          Analysis of Maximum Likelihood Estimates

                              Std        Wald
Parameter  DF  Estimate     Error      Chi-Sq    Pr > ChiSq
Intercept   1   -2.6163    0.2123    151.8266     <.0001
cat         1    0.6223    0.3193      3.7978     0.0513
age         1    0.6157    0.2838      4.7050     0.0301
ecg         1    0.3620    0.2904      1.5539     0.2126
```

```
       Partition for the Hosmer and Lemeshow Test
                          Event              Nonevent
Group  Total  Observed  Expected  Observed  Expected
1       274     17       18.66      257      255.34
2        59      7        5.60       52       53.40
3       122     15       14.53      107      107.47
4        57      9        8.94       48       48.06
5        39      9        7.85       30       31.15
6        58     14       15.41       44       42.59

        Hosmer and Lemeshow Goodness-of-Fit Test
      Chi-Square          DF          Pr > ChiSq
      0.9474              4             0.9177
```

Questions about the above output begin on the following page.

1. Is data listing described above in events trials (ET) format or in subject-specific (SS) format? Explain briefly.
2. How many covariate patterns are there for the model being fitted? Describe them.
3. Is the model being fitted a fully parameterized model? Explain briefly.
4. Is the model being fitted a saturated model? Explain briefly.
5. a. Is the deviance value of 0.9544 shown in the above output calculated using the deviance formula

   $$\mathrm{Dev}(\hat{\boldsymbol{\beta}}) = -2\ln(\hat{L}_{\mathrm{c}}/\hat{L}_{\max}),$$

   where $\hat{L}_{\mathrm{c}} = \mathrm{ML}$ for current model and $\hat{L}_{\max} = \mathrm{ML}$ for saturated model? Explain briefly.

   b. State the logit form of two logistic models that can be used to calculate the deviance value of 0.9544. Hint: One of these models is the model being fitted.

   c. How can the deviance value of 0.9544 be calculated using the difference between two log likelihood values obtained from the two models stated in part b? What are the values of these two log likelihood functions?

   d. What is actually being tested using this deviance statistic? Explain briefly.

   e. How can you justify that this deviance statistic is approximately chi-square under the null hypothesis that the fitted model has adequate fit to the data?

6. a. What can you conclude from the Hosmer–Lemeshow statistic provided in the above output about whether the model has lack of fit to the data? Explain briefly.
   b. Why does the output shown under "Partition for the Hosmer and Lemeshow Test" involve only 6 groups rather than 10 groups, and why is the degrees of freedom for the test equal to 4? Explain briefly.
   c. What two models are actually being compared by the Hosmer–Lemeshow statistic of 0.9474? Explain briefly.
   d. How can you choose between the two models described in part c?
   e. Does either of the two models described in part c perfectly fit the data? Explain briefly.

Additional questions using the same Evans County data described at the beginning of these exercises consider SAS output provided below for the following (interaction) logistic model:

$$\begin{aligned}\text{Logit P}(\mathbf{X}) = {} & \alpha + \beta_1\text{CAT} + \gamma_1\text{AGE} + \gamma_2\text{ECG} + \gamma_3\text{AGE} \times \text{ECG}\\ & + \delta_1\text{CAT} \times \text{AGE} + \delta_2\text{CAT} \times \text{ECG}\\ & + \delta_3\text{CAT} \times \text{AGE} \times \text{ECG}\end{aligned}$$

```
   Deviance and Pearson Goodness-of-Fit Statistics

Criterion     Value     DF    Value/DF     Pr > ChiSq
Deviance     0.0000     0        .             .
Pearson      0.0000     0        .             .

             Number of unique profiles: 8

                  Model Fit Statistics
                                        Intercept and
    Criterion     Intercept Only         Covariates
    −2 Log L         438.558              417.226

          Analysis of Maximum Likelihood Estimates

                                Std       Wald
Parameter DF  Estimate   Error     Chi-Sq   Pr > ChiSq
Intercept  1  −2.7158  0.2504   117.6116    <.0001
cat        1   0.7699  1.0980     0.4917    0.4832
age        1   0.7510  0.3725     4.0660    0.0438
ecg        1   0.7105  0.4741     2.2455    0.1340
catage     1  −0.00901 1.1942     0.0001    0.9940
catecg     1  −0.3050  1.3313     0.0525    0.8188
ageecg     1  −0.4321  0.7334     0.3471    0.5557
cae        1   0.0855  1.5245     0.0031    0.9553
```

```
        Partition for the Hosmer and Lemeshow Test

                        Event               Nonevent
Group  Total  Observed  Expected  Observed  Expected
1       274      17       17.00      257     257.00
2        59       7        7.00       52      52.00
3       122      15       15.00      107     107.00
4        57       9        9.00       48      48.00
5        39       9        9.00       30      30.00
6        58      14       14.00       44      44.00

        Hosmer and Lemeshow Goodness-of-Fit Test

    Chi-Square            DF             Pr > ChiSq
    0.0000                 4               1.0000
```

7. Is the model being fitted a fully parameterized model? Explain briefly.
8. Is the model being fitted a saturated model? Explain briefly.
9. a. Is the deviance value of 0.0000 shown in the above output calculated using the deviance formula

   $$\text{Dev}(\hat{\boldsymbol{\beta}}) = -2\ln(\hat{L}_c/\hat{L}_{\max}),$$

   where $\hat{L}_c$ = ML for current model and $\hat{L}_{\max}$ = ML for saturated model? Explain briefly.

   b. How can the deviance value of 0.0000 be calculated using the difference between two log likelihood functions?

   c. What is actually being tested using this deviance statistic? Explain briefly.
10. a. What can you conclude from the Hosmer–Lemeshow statistic provided in the above output about whether the interaction model has lack of fit to the data? Explain briefly.

    b. What two models are actually being compared by the Hosmer–Lemeshow statistic of 0.0000? Explain briefly.

    c. Does the interaction model perfectly fit the data? Explain briefly.

Test

The following questions and computer output consider a data from a cross-sectional study carried out at Grady Hospital in Atlanta, Georgia involving 289 adult patients seen in an emergency department whose blood cultures taken within 24 hours of admission were found to have Staph aureus infection (Rezende et al., 2002). Information was obtained on several variables, some of which were considered risk factors for methicillin-resitance (MRSA). The outcome variable is MRSA status (1 = yes, 0 = no), and covariates of interest included the following variables: PREVHOSP (1 = previous hospitalization, 0 = no previous hospitalization), AGE (continuous), GENDER (1 = male, 0 = female), and PAMU (1 = antimicrobial drug use in the previous 3 months, 0 = no previous antimicrobial drug use).

The SAS output provided below was obtained for the following logistic model:

$$\text{Logit P}(\mathbf{X}) = \alpha + \beta_1\text{PREVHOSP} + \beta_2\text{AGE} + \beta_3\text{GENDER} + \beta_4\text{PAMU}$$

```
     Deviance and Pearson Goodness-of-Fit Statistics

Criterion      Value      DF    Value/DF    Pr > ChiSq
Deviance     159.2017    181     0.8796       0.8769
Pearson      167.0810    181     0.9231       0.7630

              Number of unique profiles: 186

                   Model Fit Statistics
                                              Intercept and
Criterion         Intercept Only               Covariates
-2 Log L             387.666                    279.317

           Analysis of Maximum Likelihood Estimates

                                           Wald
Parameter  DF  Estimate  Std Error   Chi-Sq  Pr > ChiSq
Intercept  1   -5.0583   0.7643     43.8059   <.0001
PREVHOSP   1    1.4855   0.4032     13.5745    0.0002
AGE        1    0.0353   0.00920    14.7004    0.0001
gender     1    0.9329   0.3418      7.4513    0.0063
pamu       1    1.7819   0.3707     23.1113   <.0001
```

```
        Partition for the Hosmer and Lemeshow Test

                       mrsa = 1            mrsa = 0
Group  Total  Observed  Expected  Observed  Expected
1        29        1       0.99      28       28.01
2        31        5       1.95      26       29.05
3        29        2       2.85      27       26.15
4        29        5       5.73      24       23.27
5        30       10       9.98      20       20.02
6        31       12      14.93      19       16.07
7        29       16      17.23      13       11.77
8        29       20      19.42       9        9.58
9        29       22      21.57       7        7.43
10       23       21      19.36       2        3.64

      Hosmer and Lemeshow Goodness-of-Fit Test

      Chi-Square         DF          Pr > ChiSq
        7.7793           8             0.4553
```

Questions about the above output begin on the following page.

1. Is data listing used for the above analysis in events trials (ET) format or in subject-specific format? Explain briefly.
2. How many covariate patterns are there for the model being fitted? Why are there so many?
3. Is the model being fitted a fully parameterized model? Explain briefly.
4. Is the model being fitted a saturated model? Explain briefly.
5. a. Is the deviance value of 159.2017 shown in the above output calculated using the deviance formula

   $\text{Dev}(\hat{\beta}) = -2\ln(\hat{L}_{\text{c}}/\hat{L}_{\text{max}})$,

   where $\hat{L}_{\text{c}} = \text{ML}$ for current model and $\hat{L}_{\text{max}} = \text{ML}$ for saturated model? Explain briefly.

   b. The deviance value of 159.2017 is obtained by comparing log likelihood values from two logistic models, one of which is the (no-interaction) model being fitted. Describe the other logistic model, called, say, Model 2. (Hint: You should answer this question without explicitly stating the independent variables contained in Model 2.)

   c. How can the deviance value of 159.2017 be calculated using the difference between two log likelihood values obtained from the two models described in part b? What are the values of these two log likelihood functions?

   d. Why is the deviance value of 159.2017 not distributed approximately as a chi-square variable under the null hypothesis that the no-interaction model provides adequate fit?

6. a. What can you conclude from the Hosmer–Lemeshow statistic provided in the above output about whether the model has lack of fit to the data? Explain briefly.

   b. What two models are actually being compared by the Hosmer–Lemeshow statistic of 7.7793? Explain briefly.

   c. How can you choose between the two models described in part b?

   d. Does either of the two models described in part c perfectly fit the data? Explain briefly.

7. Consider the information shown in the ouput under the heading “Partition for the Hosmer and Lemeshow Test.”

   a. Briefly describe how the 10 groups shown in the output under “Partition for the Hosmer and Lemeshow Test” are formed.

   b. Why does not each of the 10 groups have the same total number of subjects?

   c. For group 5, describe how the expected number of cases (i.e., mrsa = 1) and expected number of noncases (i.e., mrsa = 0) are computed.

   d. For group 5, compute the two values that are included as two of the terms in summation formula for the Hosmer–Lemeshow statistic.

   e. How many terms are involved in the summation formula for the Hosmer–Lemeshow statistic?

Additional questions consider SAS output provided below for the following logistic model:

$$\begin{aligned}\text{Logit P}(\mathbf{X}) = {} & \alpha + \beta_1\text{PREVHOSP} + \gamma_1\text{AGE} + \gamma_2\text{GENDER} \\ & + \gamma_3\text{PAMU} + \delta_1\text{PRHAGE} + \delta_2\text{PRHGEN} \\ & + \delta_2\text{PRHPAMU}\end{aligned}$$

```
     Deviance and Pearson Goodness-of-Fit Statistics

Criterion     Value      DF    Value/DF    Pr > ChiSq
Deviance    157.1050    178     0.8826       0.8683
Pearson     159.8340    178     0.8979       0.8320

                   Model Fit Statistics

Criterion   Intercept Only   Intercept and Covariates
−2 Log L       387.666                277.221
```

```
     Partition for the Hosmer and Lemeshow Test

                      mrsa = 1            mrsa = 0
Group  Total  Observed  Expected  Observed  Expected
1        29        1       1.50       28      27.50
2        30        2       2.44       28      27.56
3        29        4       3.01       25      25.99
4        29        5       4.76       24      24.24
5        29       10       7.87       19      21.13
6        29       11      12.96       18      16.04
7        31       17      18.27       14      12.73
8        32       22      21.93       10      10.07
9        31       24      23.85        7       7.15
10       20       18      17.40        2       2.60

      Hosmer and Lemeshow Goodness-of-Fit Test

     Chi-Square           DF          Pr > ChiSq
     2.3442                8            0.9686
```

8. Is the model being fitted a fully parameterized model? Explain briefly.
9. Is the model being fitted a saturated model? Explain briefly.
10. a. Is the deviance value of 157.1050 shown in the above output calculated using the deviance formula

    $\text{Dev}(\hat{\boldsymbol{\beta}}) = -2\ln(\hat{L}_{\text{c}}/\hat{L}_{\text{max}})$,

    where $\hat{L}_{\text{c}} = \text{ML}$ for current model and $\hat{L}_{\text{max}} = \text{ML}$ for saturated model? Explain briefly.

    b. Why cannot you use this deviance statistic to test whether the interaction model provides adequate fit to the data? Explain briefly.
11. a. What can you conclude from the Hosmer–Lemeshow statistic provided in the above output about whether the interaction model has lack of fit to the data? Explain briefly.

    b. Based on the Hosmer–Lemeshow test results for both the no-interaction and interaction models, can you determine which of these two models is the better model? Explain briefly.

    c. How can you use the deviance values from the output for both the interaction and no-interaction models to carry out an LR test that compares these two models? In your answer, state the null hypothesis being tested, the formula for the LR statistic using deviances, carry out the computation of the LR test and draw a conclusion of which model is more appropriate.

Answers to Practice Exercises

1. The data listing is in events trials (ET) format. There are eight lines of data corresponding to the distinct covariate patterns defined by the model; each line contains the number of cases (i.e., events) and the number of subjects (i.e., trials) for each covariate pattern.
2. There are eight covariate patterns:

   Pattern 1: $\mathbf{X} = (\text{CAT} = 0, \text{AGE} = 0, \text{ECG} = 0)$
   Pattern 2: $\mathbf{X} = (\text{CAT} = 0, \text{AGE} = 1, \text{ECG} = 0)$
   Pattern 3: $\mathbf{X} = (\text{CAT} = 0, \text{AGE} = 0, \text{ECG} = 1)$
   Pattern 4: $\mathbf{X} = (\text{CAT} = 0, \text{AGE} = 1, \text{ECG} = 1)$
   Pattern 5: $\mathbf{X} = (\text{CAT} = 1, \text{AGE} = 0, \text{ECG} = 0)$
   Pattern 6: $\mathbf{X} = (\text{CAT} = 1, \text{AGE} = 1, \text{ECG} = 0)$
   Pattern 7: $\mathbf{X} = (\text{CAT} = 1, \text{AGE} = 0, \text{ECG} = 1)$
   Pattern 8: $\mathbf{X} = (\text{CAT} = 1, \text{AGE} = 1, \text{ECG} = 1)$
3. No. The model contains four parameters, whereas there are eight covariate patterns.
4. No. The model does not perfectly predict the case/noncase status of each of the 609 subjects in the data.
5. a. No. The deviance value of 0.9544 is not calculated using the deviance formula

      $\text{Dev}(\hat{\boldsymbol{\beta}}) = -2\ln(\hat{L}_c/\hat{L}_{\max})$.

      In particular $-2\ln\hat{L}_c = 418.181$ and $-2\ln\hat{L}_{\max} = 0$, so $\text{Dev}(\hat{\boldsymbol{\beta}}) = 418.181$.

   b. $$\begin{aligned}\text{Model 1}: \text{Logit P}(\mathbf{X}) &= \alpha + \beta\text{CAT} + \gamma_1\text{AGE} + \gamma_2\text{ECG}\\ \text{Model 2}: \text{Logit P}(\mathbf{X}) &= \alpha + \beta\text{CAT} + \gamma_1\text{AGE} + \gamma_2\text{ECG}\\ &\quad +\gamma_3\text{AGE} \times \text{ECG}\\ &\quad +\delta_1\text{CAT} \times \text{AGE}\\ &\quad +\delta_2\text{CAT} \times \text{ECG}\\ &\quad +\delta_3\text{CAT} \times \text{AGE} \times \text{ECG}\end{aligned}$$

   c. $0.9544 = -2\ln\hat{L}_{\text{Model 1}} - (-2\ln\hat{L}_{\text{Model 2}})$,
      where $-2\ln\hat{L}_{\text{Model 1}} = 418.1810$ and
      $-2\ln\hat{L}_{\text{Model 2}} = 418.1810 - 0.9544 = 417.2266$.

   d. H_0: $\delta_1 = \delta_2 = \delta_3 = 0$, i.e., the deviance is used to test for whether the coefficients of all the product terms in Model 2 are collectively nonsignificant.

   e. G = no. of covariate patterns $= 8 << n = 609$.
6. a. The HL test has a P-value of 0.9177, which is highly nonsignificant. Therefore, the HL test indicates that the model does not have lack of fit.

   b. The model contains only eight covariate patterns, so it is not possible to obtain more than eight distinct predicted risk values from the data. The degrees of freedom is 4 because it is calculated as the number of groups (i.e., 6) minus 2.

   c. Models 1 and Models 2 as stated in the answer to question 5b.

d. The deviance of 0.9544 is equivalent to the LR test that compares Model 1 with Model 2. Since this test statistic (df = 3) is highly nonsignificant, we would choose Model 1 over Model 2.

e. Neither of the two models of part c perfectly fit the data for each subject. However, since Model 2 is fully parameterized, it perfectly predicts the group proportions.

7. Yes, the interaction model is fully parameterized since the model contains eight parameters and there are eight distinct covariate patterns.

8. No, as with the no-interaction model, the interaction model does not perfectly predict the case/noncase status of each of the 609 subjects in the data.

9. a. No. The deviance value of 0.0000 is not calculated using the deviance formula

$$\text{Dev}(\hat{\beta}) = -2\ln(\hat{L}_\text{c}/\hat{L}_\text{max}).$$

In particular $-2\ln\hat{L}_\text{c} = 417.226$ and $-2\ln\hat{L}_\text{max} = 0$, so $\text{Dev}(\hat{\boldsymbol{\beta}}) = 417.226$.

b. $0.0000 = -2\ln\hat{L}_{\text{Model 2}} - (-2\ln\hat{L}_{\text{Model 2}})$. The two log likelihood functions are identical since the deviance statistic is comparing the current model (i.e., Model 2) to the fully parameterized model (i.e., Model 2).

c. What is actually being tested is whether or not Model 2 is a fully parameterized model.

10. a. The HL statistic of 0.0000 indicates that the interaction model is a fully parameterized model and therefore perfectly predicts the group proportion for each covariate pattern.

b. The same two models are being compared by the HL statistic of 0.0000, i.e., Model 2.

c. No and Yes. The interaction model does not perfectly predict each subject's response but it does perfectly predict the group proportion for each covariate pattern.

10 Assessing Discriminatory Performance of a Binary Logistic Model: ROC Curves

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_10, © Springer Science+Business Media, LLC 2010

Introduction

In this chapter, we describe and illustrate methods for assessing the extent that a fitted binary logistic model can be used to distinguish the observed cases ($Y = 1$) from the observed noncases ($Y = 0$).

One approach for assessing such discriminatory performance involves using the fitted model to predict which study subjects will be cases and which will not be cases and then determine the proportions of observed cases and noncases that are correctly predicted. These proportions are generally referred to as *sensitivity and specificity parameters*.

Another approach involves plotting a *receiver operating curve* (*ROC*) for the fitted model and computing the area under the curve as a measure of discriminatory performance. The use of ROCs has become popular in recent years because of the availability of computer software to conveniently produce such a curve as well as compute the area under the curve.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

Objectives

Upon completing this chapter, the learner should be able to:

1. Given a fitted binary logistic model, describe or illustrate how a cut-point can be used to classify subjects as predicted cases ($Y = 1$) and predicted noncases ($Y = 0$).
2. Given a fitted binary logistic model, describe or illustrate how a cut-point can be used to form a misclassification (or diagnostic table).
3. Define and illustrate what is meant by true positives, false positives, true negatives, and false negatives.
4. Define and illustrate what is meant by sensitivity and specificity.
5. Define and illustrate "perfect discrimination."
6. Describe what happens to sensitivity and specificity parameters when a cut-point used for discrimination of a fitted logistic model decreases from 1 to 0.
7. Describe what happens to (1 − specificity) when a cut-point used for discrimination decreases from 1 to 0.
8. State one or more uses of an ROC curve.
9. State and/or describe briefly how an ROC curve is constructed.
10. State and/or describe briefly how the area under an ROC curves is calculated.
11. Describe briefly how to interpret a calculated area under an ROC curve in terms of the discriminatory performance of a fitted logistic model.
12. Given a printout of a fitted binary logistic model, evaluate how well the model discriminates cases from noncases.

Presentation

I. Overview

Focus Assessing discriminatory performance (**DP**) of a binary logistic model

Good DP: model discriminates **cases** $(Y=1)$ from **noncases** $(Y=0)$

This presentation describes how to assess *discriminatory performance* (*DP*) of a binary logistic model.

We say that a model provides *good DP* if the covariates in the model help to predict (i.e., discriminate) which subjects will develop the outcome (Y = 1, or the *cases*) and which will not develop the outcome (Y = 0, or the *noncases*).

Example: Blunt knee trauma ⇒ X-ray?
Predictor variables:
ability to flex knee
ability to put weight on knee
patient's age
injury to knee head
injury to patella
Outcome variable:
knee fracture status

For example, we may wish to determine whether or not a subject with blunt knee trauma should be sent for an X ray based on a physical exam that measures ability to flex knee, ability to put weight on knee, injury to knee head, injury to patella, and age. The outcome here is whether or not the person has a knee fracture.

Approach 1
Use fitted model to predict which subjects will be cases or noncases e.g.,

If $\hat{\mathrm{P}}(\mathbf{X}) > 0.2$, predict subj **X** to be case,
if $\hat{\mathrm{P}}(\mathbf{X}) \leq 0.2$, predict subj **X** to be noncase, where **cut-point** = 0.2

Note: Rare outcome ⇒ 0.2, or even 0.02, high

One way to measure *DP* involves using the fitted model to decide how to predict which subjects will be cases and which will be noncases. For example, one may decide that if the predicted probability for subject **X** (i.e., $\hat{\mathrm{P}}(\mathbf{X})$) is greater than 0.2, we will predict that subject **X** will be a case, whereas otherwise, a noncase. The value of 0.2 used here is called a *cut-point*. Note that for a very rare health outcome, a predicted probability of 0.2, or even 0.02, could be considered a high "risk."

Classification/Diagnostic Table

True (Observed) Outcome

		Y = 1	Y = 0
Predicted	Y = 1	$n_{\mathrm{TP}} = 70$	20
Outcome	Y = 0	30	$n_{\mathrm{TN}} = 80$
		$n_1 = 100$	$n_0 = 100$

$n_{\mathrm{TP}} = \#$ of true +,
$n_{\mathrm{TN}} = \#$ of true −,

The observed and predicted outcomes are combined into a *classification or diagnostic table*, an example of which is shown at the left, In this table, we focus on two quantities: the number of true cases (i.e., we are assuming that the observed cases are the true cases) that are predicted to be cases (*true positives or TP*), and the number of true noncases that are predicted to be noncases (*true negatives or TN*).

$\mathbf{Se} = n_{\text{TP}}/n_1 = 70/100 = 0.7$

$\mathbf{Sp} = n_{\text{TN}}/n_0 = 80/100 = 0.8$

Perfect (ideal) discrimination:

Se = Sp = 1

The proportion of true positives among all cases is called *sensitivity* **(Se)**, and the proportion of true negatives among all noncases is called the *specificity* **(Sp)**. Ideally, perfect discrimination would occur if *both sensitivity and specificity are equal to 1.*

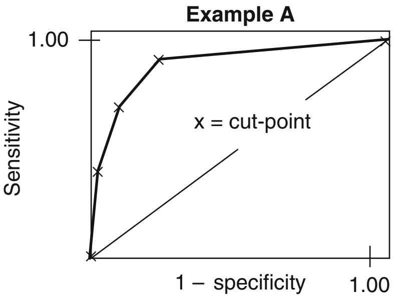
Example: cut-point = 0.2:
Model 1: Se = 0.7 and Sp = 0.8
better **DP** than
Model 2: Se = 0.6 and Sp = 0.5

Thus, *for a given cut-point, the closer both the sensitivity and specificity are to 1, the better the discriminatory performance* (see example at left).

Drawback: Sensitivity and specificity varies by cut-point

ROC curve:
considers Se and Sp for a range of cut-points.

A drawback to measuring discrimination as described above is that the sensitivity and specificity that results from a given cut-point may vary with the cut-point chosen. An alternative approach involves obtaining a summary measure based on a range of cut-points chosen for a given model. Such a measure is available from an *ROC curve*.



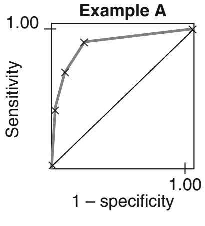
ROC stands for *receiver operating characteristic*, which was originally developed in the context of electronic signal detection. When applied to a logistic model, an ROC is a plot of *sensitivity* **(Se)** vs. *1 – specificity* **(1 – Sp)** *derived from several cut-points for the predicted value*.

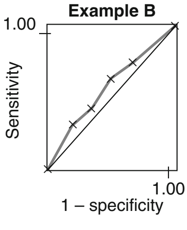
$$\mathbf{1 - Sp} = \frac{\text{falsely predicted cases}}{\text{observed noncases}} = \frac{n_{\text{FP}}}{n_0}$$

Want:

1 – Sp close to 0 and **Se > 1 – Sp**

Note that **1 – Sp** gives the proportion of observed noncases that are (falsely) predicted to be cases, i.e., 1 – Sp gives the proportion of *false positives* **(FPs)**. Since we want both Se and Sp close to 1, we would like **1 – Sp** *close to zero*, and moreover, we would *expect* **Se** *to be larger than* **1 – Sp**, as in the above graph.





Key: The larger the area under the curve, the better is the **DP**.

ROC curves for two different models based on the same data are shown at the left. These graphs may be compared according to the following criterion: *The larger the area under the curve, the better is the discrimination.* In our example, we see that the area in Example A is larger than the area in Example B, indicating that the model used in Example A discriminates better than the model in Example B.

Why does area under ROC measure DP? See Section III.

Why does the area under the ROC measure discriminatory performance **(DP)**? We discuss this question and other characteristics of ROC curves in **Section III** of this chapter.

II. Assessing Discriminatory Performance Using Sensitivity and Specificity Parameters

Cut-point can be used with $\hat{\mathrm{P}}(\mathbf{X})$ to predict whether subject is case or noncase.

In the previous section, we illustrated how a *cut-point* could be used with a fitted logistic model to assign a subject **X** based on the predicted value $\hat{\mathbf{P}}(\mathbf{X})$ to be a "predicted" case or noncase.

If $\hat{\mathrm{P}}(\mathbf{X}) > c_{\mathrm{p}}$, predict subj **X** to be case.
If $\hat{\mathrm{P}}(\mathbf{X}) \leq c_{\mathrm{p}}$, predict subj **X** to be noncase.

Denoting the general cut-point as $\mathbf{c_p}$, we typically predict a subject to be a case if $\hat{\mathrm{P}}(\mathbf{X})$ exceeds $\mathbf{c_p}$ vs. a noncase if $\hat{\mathrm{P}}(\mathbf{X})$ doesn not exceed $\mathbf{c_p}$.

Table 10.1
General Classification/Diagnostic Table

		True (Observed) Outcome	
	$\mathbf{c_p}$	$Y = 1$ (case)	$Y = 0$ (noncase)
Predicted	$Y = 1$	n_{TP}	n_{FP}
Outcome	$Y = 0$	n_{FN}	n_{TN}
		n_1	n_0

Given a cut-point $\mathbf{c_p}$, the observed and predicted outcomes can then be combined into a *classification* (*diagnostic*) *table*, the general form of which is shown here. The cell frequencies within this table give the number of *true positives* (n_{TP}) and *false negatives* (n_{FN}) out of the number of *true cases* (n_1), and the number of *false positives* (n_{FP}) and *true negatives* (n_{TN}) out of the number of true noncases (n_0).

$$\begin{aligned}\mathbf{Se} &= \Pr(\text{true positive} \mid \text{true case}) \\ &= n_{\mathrm{TP}}/n_1 \\ \mathbf{Sp} &= \Pr(\text{true negative} \mid \text{true noncase}) \\ &= n_{\mathrm{TN}}/n_0\end{aligned}$$

From the classification table, we can compute the *sensitivity* **(Se)** and the *specificity* **(Sp)**.

Perfect Discrimination (Se = Sp = 1)

		True (Observed) Outcome	
	$\mathbf{c_p}$	$Y = 1$	$Y = 0$
Predicted	$Y = 1$	n_{TP}	0
Outcome	$Y = 0$	0	n_{TN}
		n_1	n_0

Ideally, perfect discrimination would occur if *both sensitivity and specificity are equal to 1*, which would occur if there were no false negatives ($n_{\mathrm{FN}} = 0$) and no false positives ($n_{\mathrm{FP}} = 0$).

Sp and Se values vary with $\mathbf{c_p}$

In our overview, we pointed out that the sensitivity and specificity values that result from a given cut-point may vary with the cut-point chosen.

Two different $\mathbf{c_p}$s: $\mathbf{c_p = 1}$ and $\mathbf{c_p = 0}$

As a simple illustration, suppose the following two extreme cut-points are used: $c_p = 1$ and $c_p = 0$. The corresponding classification tables for each of these cut-points are shown below at the left.

$\mathbf{c_p = 1}$: Se = 0, Sp = 1

OBS Y

		$Y = 1$	$Y = 0$
PRED	$Y = 1$	0	0
Y	$Y = 0$	n_{FN}	n_{TN}
		n_1	n_0

If the cut point is $c_p = 1$, then assuming that $\hat{P}(\mathbf{X}) = 1$ is not attained for any subject, there will be *no predicted cases* among either the n_1 true cases or the n_0 true noncases. For this situation, then, the sensitivity is 0 and the specificity is 1.

$\mathbf{c_p = 0}$: Se = 1, Sp = 0

OBS Y

		$Y = 1$	$Y = 0$
PRED	$Y = 1$	n_{TP}	n_{FP}
Y	$Y = 0$	0	0
		n_1	n_0

On the other hand, if the cut-point is $c_p = 0$, then assuming that $\hat{P}(\mathbf{X}) = 0$ is not attained for any subject, there will be *no predicted noncases* among either the n_1 true cases or the n_0 true noncases. For this situation, then, the sensitivity is 1 and the specificity is 0.

Question: c_p decreases from 1 to 0?

Answer: Se increases from 0 to 1
Sp decreases from 1 to 0

Let us now consider what would happen if c_p decreases from 1 to 0. As we will show by example, as c_p decreases from 1 to 0, the sensitivity will increase from 0 to 1 whereas the specifity will decrease from 1 to 0.

EXAMPLE

Table 10.2
Classification Tables for Two Models by Varying Classification Cut-Point (c_p)

MODEL 1 | MODEL 2

$c_p = 1.00$

Model 1: Se = 0.00, Sp = 1.00

PRED Y	OBS Y: Y = 1	Y = 0
Y = 1	0	0
Y = 0	100	100

Model 2: Se = 0.00, Sp = 1.00

PRED Y	OBS Y: Y = 1	Y = 0
Y = 1	0	0
Y = 0	100	100

$c_p = 0.75$

Model 1: Se = 0.10, Sp = 1.00

PRED Y	OBS Y: Y = 1	Y = 0
Y = 1	10	0
Y = 0	90	100

Model 2: Se = 0.10, Sp = 0.90

PRED Y	OBS Y: Y = 1	Y = 0
Y = 1	10	0
Y = 0	90	90

$c_p = 0.50$

Model 1: Se = 0.60, Sp = 1.00

PRED Y	OBS Y: Y = 1	Y = 0
Y = 1	60	0
Y = 0	40	100

Model 2: Se = 0.60, Sp = 0.40

PRED Y	OBS Y: Y = 1	Y = 0
Y = 1	60	60
Y = 0	40	40

$c_p = 0.25$

Model 1: Se = 1.00, Sp = 1.00

PRED Y	OBS Y: Y = 1	Y = 0
Y = 1	100	0
Y = 0	0	100

Model 2: Se = 0.80, Sp = 0.20

PRED Y	OBS Y: Y = 1	Y = 0
Y = 1	80	80
Y = 0	20	20

$c_p = 0.10$

Model 1: Se = 1.00, Sp = 0.40

PRED Y	OBS Y: Y = 1	Y = 0
Y = 1	100	60
Y = 0	0	40

Model 2: Se = 0.90, Sp = 0.10

PRED Y	OBS Y: Y = 1	Y = 0
Y = 1	90	90
Y = 0	10	10

$c_p = 0.00$

Model 1: Se = 1.00, Sp = 0.00

PRED Y	OBS Y: Y = 1	Y = 0
Y = 1	100	100
Y = 0	0	0

Model 2: Se = 1.00, Sp = 0.00

PRED Y	OBS Y: Y = 1	Y = 0
Y = 1	100	100
Y = 0	0	0

We illustrate on the left the classification tables and corresponding sensitivity and specificity values obtained from varying the cut-points for two hypothetical logistic regression models.

Based on this information, what can you conclude *for each model* separately as to how the sensitivity changes as the cut-point $\mathbf{c_p}$ decreases from 1.00 to 0.75 to 0.50 to 0.25 to 0.10 to 0.00? Similarly, what can you conclude for each model as to how the specificity changes as the cut-point decreases from 1.00 to 0.00?

The answers to the above two questions are that *for both models, as the cut-put c_p decreases* from 1.00 to 0.00, *the sensitivity increases* from 0.00 to 1.00 and *the specificity decreases* from 1.00 to zero. Note that this result will always be the case for *any* binary logistic model.

Next question: For each model separately, as the cut-point decreases, *does the sensitivity increase at a faster rate than the specificity decreases*?

The answer to the latter question depends on which model we consider. For Model 1, the answer is *yes*, since the sensitivity starts to change immediately as the cut-point changes, whereas the specificity remains at 1 until the cut-point changes to 0.10.

For Model 2, however, the answer is *no*, because the sensitivity increases at the same rate that the specificity decreases. In particular, the sensitivity increases by 0.10 (from 0.00 to 0.10) while the sensitivity decreases by 0.10 (from 1.00 to 0.90), followed by correspondingly equal changes of 0.50, 0.20, 0.10 and 0.10 as the cut-point decreases to 0.

Sp may change at a different rate than **Se**

So, even though the sensitivity increases and the specificity decreases as the cut-point decreases, *the specificity may change at a different rate than the sensitivity* depending on the model being considered.

EXAMPLE (continued)

Table 10.3
Summary of Classification Information For Models 1 and 2 (incl. 1 − Specificity)

MODEL 1:

c_p	1.00	0.75	0.50	0.25	0.10	0.00
Se	**0.00**	**0.10**	**0.60**	**1.00**	**1.00**	**1.00**
Sp	1.00	1.00	1.00	1.00	0.40	0.00
1 – Sp	**0.00**	**0.00**	**0.00**	**0.00**	**0.60**	**1.00**

MODEL 2:

c_p	1.00	0.75	0.50	0.25	0.10	0.00
Se	**0.00**	**0.10**	**0.60**	**0.80**	**0.90**	**1.00**
Sp	1.00	0.90	0.40	0.20	0.10	0.00
1 – Sp	**0.00**	**0.10**	**0.60**	**0.80**	**0.90**	**1.00**

An alternative way to evaluate the discrimination performance exhibited in a classification table is to consider "1 – specificity" (**1 – Sp**) instead of "specificity" in addition to the sensitivity.

The tables at the left summarize the results of the previous misclassification tables, and they include **1 – Sp** values as additional summary information.

Model 1: Se increases at faster rate than 1 − Sp

For Model 1, when we **compare Se to 1 – Sp** values as the cut-point decreases, we see that the Se values increase at a faster rate than the values of 1 − Sp.

Model 2: Se and 1 − Sp increase at same rate

For Model 2, however, we find that both Se and 1 − Sp values increase at the exact same rate.

1 – Sp more appealing than Sp because
Se and 1 – Sp both focus on predicted cases

Using 1 – Sp instead of Sp is descriptively appealing for the following reason: both Se and 1 – Sp focus, respectively, on the probability of being either correctly or falsely predicted to be a case.

$$\mathbf{Se} = \mathbf{Prop.\ True\ Positives\ (TP)}$$
$$= n_{\mathrm{TP}}/n_1$$
where n_{TP} = correctly predicted cases

$$\mathbf{1 - Sp} = \mathbf{Prop.\ False\ Positive\ (FP)}$$
$$= n_{\mathrm{FP}}/n_0$$
where n_{FP} = falsely predicted cases

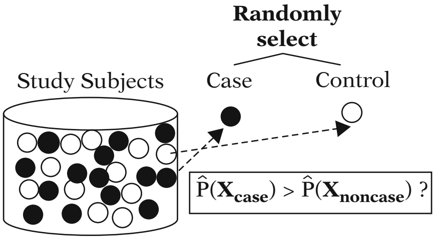
Among the observed (i.e., true) cases, Se considers the proportion of subjects who are "true positives" (TP), that is, correctly predicted as cases. Among the observed (i.e., true) noncases, 1 – Sp considers the proportion of subjects who are "false positives" (FP), that is, are falsely predicted as cases.

Good discrimination

$$\Downarrow\text{(expect)}$$

$$\mathbf{Se} = \underbrace{n_{\mathrm{TP}}}_{\text{Correctly predicted cases}}/n_1 > 1 - \mathbf{Sp} = \underbrace{n_{\mathrm{FP}}}_{\text{falsely predicted noncases}}/n_0$$

One would expect for a model that has good discrimination that the proportion of true cases that are (correctly) predicted as cases (i.e., Se) would be higher than the proportion of true noncases that are (falsely) diagnosed as cases (i.e., 1 − Sp). Thus, to evaluate discrimination performance, it makes sense to compare Se (i.e., involving correctly diagnosed cases) with 1 − Sp (i.e., involving falsely predicted noncases).


Randomly select
Study Subjects
Case
Control
$\hat{P}(\mathbf{X}_{\text{case}}) > \hat{P}(\mathbf{X}_{\text{noncase}})$?

EXAMPLE

Table 10.3: Collectively compare Se with 1 − Sp over all cut-points

MODEL 1:

c_p	1.00	0.75	0.50	0.25	0.10	0.00
Se	0.00	0.10	0.60	1.00	1.00	1.00
>?	**No**	**Yes**	**Yes**	**Yes**	**Yes**	**No**
1 – Sp	0.00	0.00	0.00	0.00	0.60	1.00

Good discrimination: **Se** > **1 – Sp** overall

MODEL 2:

c_p	1.00	0.75	0.50	0.25	0.10	0.00
Se	0.00	0.10	0.60	0.80	0.90	1.00
>?	**No**	**No**	**No**	**No**	**No**	**No**
1 – Sp	0.00	0.10	0.60	0.80	0.90	1.00

Poor discrimination: **Se** never > **1 – Sp**
(**Here: Se** = **1** − **Sp always**)

Problem with using above info:
Se and 1 − Sp values are **summary statistics** for several subjects based on a specific cut-point

Better approach:
Compute and compare **predicted probabilities for specific pairs of subjects**

⇓

Obtained via **ROC curves**
(next section)

Returning to Table 10.3, *suppose we pick a case and a noncase at random from the subjects analyzed in each model. Is the case or the noncase more likely to have a higher predicted probability?*

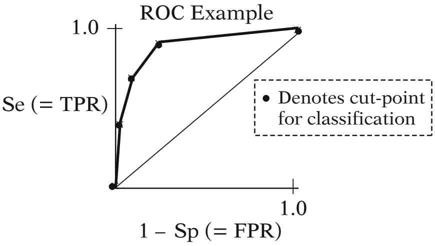
Using Table 10.3, we can address this question by "collectively" comparing for each model, the proportion of true positives (Se) with the corresponding proportion of false positives (1 − Sp) over all cut-points considered.

For Model 1, we find that at each cut-point, the proportion of true positives is larger than the proportion of false positives at each cut-point except when $c_p = 1.00$ or 0.00, at which both proportions are equal. These results suggest that Model 1 provides good discrimination since, overall, Se values are greater than 1 − Sp values.

For Model 2, however, we find that at each cut-point, the proportion of true positives is identical to the proportion of false positives at each cut-point. These results suggest that Model 2 does not provide good discrimination, since Se is never greater (although also never less) than 1 − Sp.

Nevertheless, the use of information from Table 10.3 is not the best way to compare predicted probabilities obtained from randomly selecting a case and noncase from the data. The reason: sensitivity and 1 – specificity values are *summary statistics* for several subjects based on a specific cut-point; what is needed instead is to compute and compare *predicted probabilities for specific pairs of subjects*. The use of ROC curves, which we describe in the next section, provides an appropriate way to quantify and compare such predicted probabilities.

III. Receiver Operating Characteristic (ROC) Curves



A Receiver Operating Curve (ROC) is a plot of **sensitivity (Se) by 1 – specificity (1 – Sp)** values derived from several classification tables corresponding to different cut-points used to classify subjects into one of two-groups, e.g., predicted cases and noncases of a disease.

Equivalently, the ROC is a plot of the *true positive rate* (TPR = Se) *by the false positive rate* (*FPR* = 1 − Sp).

ROC history:

- Developed by engineers in WW II to detect enemy objects (signal detection), i.e., $\hat{P}(\mathbf{X})$ is a radar signal
- Now used in medicine, radiology, psychology, machine learning, data mining

As described in Wikipedia (a free Web-based encyclopedia), "the ROC was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battle fields, also known as the signal detection theory; in this situation, a signal represents the predicted probability that a given object is an enemy weapon." ROC analysis is now widely used in medicine, radiology, psychology and, more recently in the areas of machine learning and data mining.

ROC from logistic model:

- Assesses overall how well model predicts who will or will not have the outcome
- Measures how well fitted model discriminates true (observed) cases from true (observed) noncases

Equivalent

When using an ROC derived from a logistic model used to predict a binary outcome, the ROC allows for an overall assessment of how well the model predicts who will have the outcome and who will not have the outcome. Stated another way in the context of epidemiologic research, the ROC provides a measure of how well the fitted model distinguishes true cases (i.e., those observed to have the outcome) from true noncases (i.e., those observed not to have the outcome).

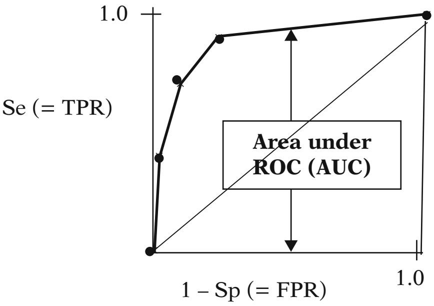
ROC provides answer to:

If $\mathbf{X}_{\textbf{true case}}$ and $\mathbf{X}_{\textbf{true noncase}}$ are covariate values for a randomly chosen case/noncase pair,

will

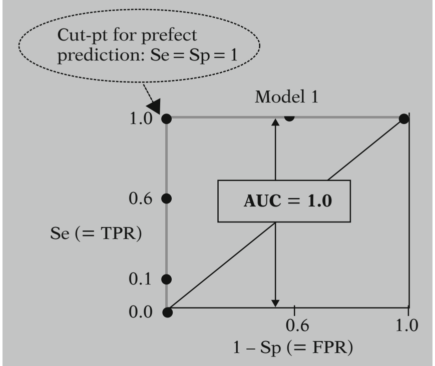
$$\hat{P}(\mathbf{X}_{\textbf{true case}}) > \hat{P}(\mathbf{X}_{\textbf{true noncase}})?$$

More specifically, an ROC provides an appropriate answer to the question we previously asked when we compared classification tables for two models: How often will a randomly chosen (true) case have a higher probability of being predicted to be a case than a randomly chosen true noncase?



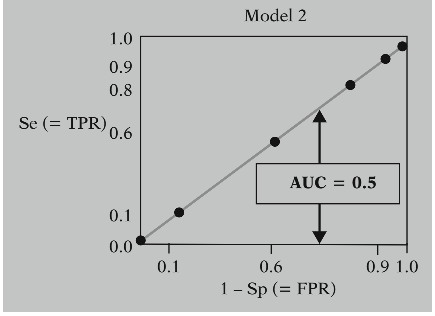
Moreover, we will see that the answer to this question can be quantified by obtaining the area under an ROC curve (AUC): the larger the area, the better the discrimination.

EXAMPLE



First, we provide the two ROCs derived from hypothetical Models 1 and 2 that we considered in the previous section. Notice that the ROC for each model is determined by connecting the dots that plot pairs of Se and 1 − Sp values obtained for several classification cut-points.

For Model 1, the area under the ROC is 1.0.



In contrast, for Model 2, the area under the ROC is 0.5.

Since the area under the ROC for Model 1 is twice that for Model 2, we would conclude that *Model 1 has better discriminatory performance than Model 2.*

So why is Model 1 a better discriminator than Model 2?

Good discrimination
$$\Leftrightarrow$$
$$\text{TPR} > \text{FPR}$$
where
$$\text{Se} = \text{TPR},\ 1 - \text{Sp} = \text{FPR}$$

How can we explain this conceptually?
Our explanation:
The AUC measures *discrimination*, that is, the ability of the model to correctly classify those with and without the disease. We would expect a model that provides good discrimination to have the property that true cases have a higher predicted probability (of being classified as a case) than true noncases. In other words, we would expect the true positive rate (TPR = Se) to be higher than the false positive rate (FPR = 1 − Sp) for all cut-points.

Model 1: TPR $\geq$ FPR always
$\Downarrow$
Excellent discrimination

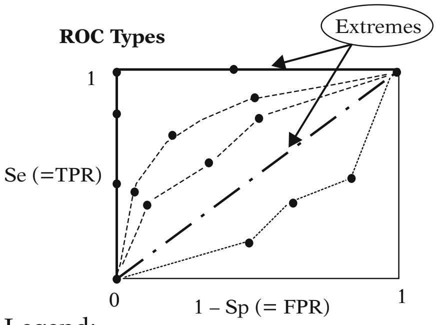
Observing the above ROCs, we see that, for Model 1, TPR (i.e., Se) is consistently higher than its corresponding FPR (i.e., 1 − Sp); so, this indicates that Model 1 does well in differentiating the true cases from the true noncases.

Model 2: TPR = FPR always
$\Downarrow$
No discrimination

In contrast, for Model 2 corresponding true positive and false positive rates are always equal, which indicates that Model 2 fails to differentiate true cases from true noncases.

Two extremes:
Model 1: perfect discrimination
Model 2: no discrimination

The two ROCs we have shown actually represent two extremes of what typically results for such plots. Model 1 gives perfect discrimination whereas Model 2 gives no discrimination.



Legend:
- ——— perfect discrimination (Area = 1.0)
- ------------ positive discrimination (0.5 < Area ≤ 1.0)
- .... negative discrimination (0.0 ≤ Area < 0.5)
- — · — no discrimination (Area = 0.5)

We show in the figure at the left several different types of ROCs that may occur. Typically, as shown by the two dashed curves, the ROC plot will lie above the central diagonal (45°) line that corresponds to Se = 1 − Sp; for such curves, the AUC is at least 0.5.

It is also possible that the ROC may lie completely below the diagonal line, as shown by the dotted curve near the bottom of the figure, in which case the AUC is less than 0.5. This situation indicates negative discrimination, i.e., the model predicts true noncases better (i.e., higher predicted probability) than it predicts true cases.

An AUC of exactly 0.5 indicates that the model provides no discrimination, i.e., predicting the case/noncase status of a randomly selected subject is equivalent to flipping a fair coin.

Grading Guidelines for AUC values:

0.90–1.0 = excellent discrimination (A)
0.80–0.90 = good discrimination (B)
0.70–0.80 = fair discrimination (C)
0.60–0.70 = poor discrimination (D)
0.50–0.60 = failed discrimination (F)

A rough guide for grading the discriminatory performance indicated by the AUC follows the traditional academic point system, as shown on the left.

However:

- Unusual to find AUC ≥ 0.9
- If so, there is nearly *complete separation of data points*

$\Downarrow$

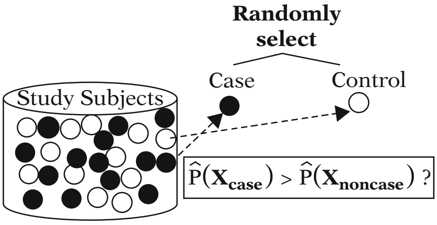
	E	Not E	
D	n_1	0	
Not D	0	n_0	$\widehat{\text{OR}}$ undefined
	n_1	n_0	

Note, however, that it is typically unusual to obtain an AUC as high as 0.90, and if so, almost all exposed subjects are cases and almost all unexposed subjects are noncases (i.e., there is nearly complete separation of data points). When there is such "complete separation," it is impossible as well as unnecessary to fit a logistic model to the data.

IV. Computing the Area Under the ROC (AUC)

In this section, we return to the previously asked question:

Suppose we pick a case and a noncase at random from the subjects analyzed using a logistic regression model. Is the case or the noncase more likely to have a higher predicted probability?


Randomly select
Case
Control
Study Subjects
$\hat{\text{P}}(\mathbf{X}_{\mathbf{case}}) > \hat{\text{P}}(\mathbf{X}_{\mathbf{noncase}})$?

$$p_{\text{d}} = \frac{\text{no. of pairs in which } \hat{\text{P}}(\mathbf{X}_{\mathbf{case}}) \geq \hat{\text{P}}(\mathbf{X}_{\mathbf{noncase}})}{\text{Total \# case-control pairs}}$$

To answer this question precisely, we must use the fitted model to compute the proportion of total case/noncase pairs for which the predicted value for cases is at least as large as the predicted value for noncases.

$p_{\text{d}} > 0.5 \Rightarrow \hat{\text{P}}(\mathbf{X}_{\mathbf{case}}) > \hat{\text{P}}(\mathbf{X}_{\mathbf{noncase}})$ for randomly chosen case-control pair

(expect this result if model discriminates cases from noncases)

If this proportion is larger than 0.5, then the answer is that the randomly chosen case will likely have a higher predicted probability than the randomly chosen noncase. Note that this is what we would expect to occur if the model provides at least minimal predictive power to discriminate cases from noncases.

More important:

$p_{\text{d}} = \text{AUC}$

Moreover, the actual value of this proportion, tells us much more, namely this proportion gives the "Area under the ROC" (i.e., AUC), which, as discussed in the previous section, provides an overall measure of the model's ability to discriminate cases from noncases.

EXAMPLE

Example of AUC calculation:

$n = 300$ subjects
$n_1 = 100$ true cases
$n_0 = 200$ true noncases

EXAMPLE

To illustrate the calculation of this proportion, suppose there are 300 (i.e., n) subjects in the entire study, of which 100 (i.e., n_1) are true cases and 200 (i.e., n_0) are true noncases.

Example: (Continued)

Fit logistic model P(**X**)
and
compute $\hat{\text{P}}(\mathbf{X}_i)$ for $i = 1, \ldots, 300$

$n_{\text{p}} = n_1 \times n_0 = 100 \times 200 = 20{,}000$

We then fit a logistic model P(**X**) to this data set, and we compute the predicted probability of being a case, i.e., $\hat{\text{P}}(\mathbf{X_i})$, for each of the 300 subjects.

For this dataset, the *total number of possible case/noncase pairs* (i.e., n_{p}) is the product 100 × 200, or 20,000.

w = no. of case/noncase pairs for which
$\hat{\text{P}}(\mathbf{X_{case}}) > \hat{\text{P}}(\mathbf{X_{noncase}})$

EXAMPLE

Example: Suppose $w = 11{,}480$
(i.e., 57.4% of 20,000)

We now let w denote the number of these pairs for which $\hat{\text{P}}(\mathbf{X})$ for the case is larger than $\hat{\text{P}}(\mathbf{X})$ for the corresponding control. Suppose, for example, that $w = 11{,}480$, which means that in 57.4% of the 20,000 pairs, the case had a higher predicted probability than its noncase pair.

Z = no. of case/noncase pairs for which
$\hat{\text{P}}(\mathbf{X_{case}}) = \hat{\text{P}}(\mathbf{X_{noncase}})$

EXAMPLE

Example: Suppose $z = 5{,}420$.
(i.e., 27.1% of 20,000)

$$p_{\text{d}} = \frac{w+z}{n_{\text{p}}} = \frac{11{,}480 + 5{,}420}{20{,}000} = 0.8450$$

Modified formula:

$$c = \frac{11{,}480 + 0.5(5{,}420)}{20{,}000} = \frac{14{,}190}{20{,}000} = \boxed{0.7095}$$

$$\boxed{c = \frac{w+0.5z}{n_{\text{p}}} = \text{AUC}}$$

Now let z denote the number of case/noncase pairs in which both case and noncase had exactly the same predicted probability. Continuing our example, we suppose $z = 5{,}420$, so that this result occurred for only 27.1% of the 20,000 pairs.

Then, for our example, the proportion of the 20,000 case-control pairs for which the case has at least as large a predicted probability as the control is $(w + z)/n_{\text{p}}$, which is 16,900/20,000, or 0.8450.

A modification of this formula (called "c") involves weighting by 0.5 any pair with equal predicted probabilities; that is, the numerator is modified to "$w + 0.5z$", so that c becomes 0.7095.

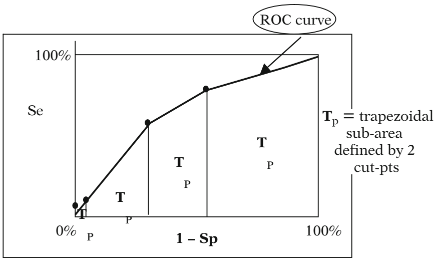
It is the latter modified formula that is equivalent to the area under the ROC, i.e., AUC.

Interpretation from guidelines:
AUC = 0.7095 ⇒ Fair discrimination
(grade C)

Based on the grading guidelines for AUC that we provided in the previous section, the AUC of 0.7095 computed for this hypothetical example would be considered to provide fair discrimination (i.e., grade C).

How does AUC formula provide geometrical area under curve?
Ilustrative Example below.

In our presentation of the above AUC formula, we have not explicitly demonstrated why this formula actually works to provide the area under the ROC curve.



We now illustrate how this numerical formula translates into the geometrical area under the curve.

The method we illustrate is often referred to as the *trapezoid method*; this is because the area directly under the curve requires the computation and summation of several trapezoidal sub-areas, as shown in the sketch at the left.

EXAMPLE

$$\hat{\mathrm{P}}(\mathbf{X}) = \frac{1}{1+\exp[-(\hat{\alpha}+\hat{\beta}_1\mathrm{X}_1+\hat{\beta}_2\mathrm{X}_2)]}$$

$$\hat{\beta}_1 > 0,\ \hat{\beta}_2 > 0$$

As in our previous example, we consider 100 cases and 200 noncases and the fitted logistic regression model shown at the left involving two binary predictors, in which both $\hat{\beta}_1$ and $\hat{\beta}_2$ are positive.

Classification information for different cut points ($\mathbf{c_p}$)

	X_1	X_2	$\hat{P}(X)$	C	NC	C+	NC+	Se%	1 – Sp%
	-	-	c_0=1	0	0	0	0	0	0
Covariate patterns	1	1	c_1	10	2	10	2	10	1
	1	0	c_2	50	48	60	50	60	25
	0	1	c_3	20	50	80	100	80	50
	0	0	c_4	20	100	100	200	100	100

c_0=1 ⇒ 0 cases (C) and
0 non-cases (NC)
test +

A classification table for these data shows four covariate patterns of the predictors that define exactly four cut points for classifying a subject as positive or negative in the construction of an ROC curve. A fifth cut point is included ($c_0 = 1$) for which nobody tests positive since $\hat{\mathrm{P}}(\mathbf{X}) \leq 1$ always. The ROC curve will be determined from a plot of these five points.

$$c_1 = \hat{\mathrm{P}}(\mathbf{X}_{c_1}) > c_2 = \hat{\mathrm{P}}(\mathbf{X}_{c_2}) > \cdots > c_4 = \hat{\mathrm{P}}(\mathbf{X}_{c_4})$$

Note that the cut points are listed in decreasing order.

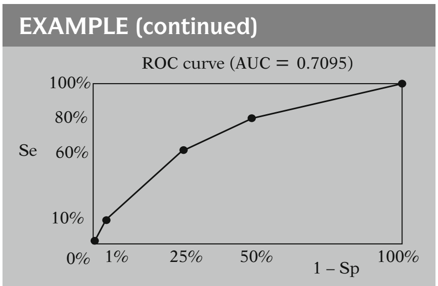
$$c_p \downarrow \ \Rightarrow \ \text{Se and } 1-\text{Sp} \uparrow$$

Also, as the cut point lowers, both the sensitivity and 1 − specificity will increase.

	Se	1 – Sp
c_1	10% cases test +	1% noncases test +
c_2	60% cases test +	25% noncases test +
c_3	80% cases test +	50% noncases test +
c_4	100% cases test +	100% noncases test +

More specifically, at cutpoint c_1, 10 of 100 cases (10%) test positive and 2 out of 200 (1%) noncases test positive. At cutpoint c_2, 60% of the cases and 25% of the noncases test positive. At cutpoint c_3, 80% of the cases and 50% of the noncases test positive. At cutpoint c_4, all 100 cases and 200 noncases test positive because $\hat{\mathrm{P}}(\mathbf{X})$ is equal to the cut point even for subjects without any risk factor ($X_1 = 0$ and $X_2 = 0$).

EXAMPLE (continued)



The resulting ROC curve is shown at the left. The AUC for this curve is 0.7095. We now show the calculation of this area using the AUC formula given earlier and using the trapezoid approach.

n_p = 100 cases × 200 noncases
= 20,000 case/noncase pairs

We first apply the above AUC formula. In our example, there are 100 cases and 200 noncases yielding 100 × 200 = 20,000 total pairs.

When $X_1 = 1$ and $X_2 = 1$:

10 cases and 2 noncases have same $\hat{\text{P}}(\mathbf{X})$,
i.e., 10 × 2 = **20 tied pairs**

10 cases have higher $\hat{\text{P}}(\mathbf{X})$ than
48 + 50 + 100 = 198 noncases
i.e., 10 × 198 = **1,980 concordant pairs**

The ten cases with $X_1 = 1$ and $X_2 = 1$ have the same predicted probability (*tied*) as the two noncases who also have $X_1 = 1$ and $X_2 = 1$.

But those same ten cases have a higher predicted probability (*concordant*) than the other 48 + 50 + 100 noncases.

When $X_1 = 1$ and $X_2 = 0$:

50 cases have lower $\hat{\text{P}}(\mathbf{X})$ than 2 noncases
i.e., 50 × 2 = **100 discordant** pairs

50 cases and 48 noncases have same $\hat{\text{P}}(\mathbf{X})$,
i.e., 50 × 48 = **2,400 tied pairs**

50 cases have higher $\hat{\text{P}}(\mathbf{X})$ than
50 + 100 = 150 noncases
i.e., 50 × 150 = **7,500 concordant** pairs

Similarly, the 50 cases with $X_1 = 1$ and $X_2 = 0$ are

discordant with the 2 noncases that have a higher predicted probability,

tied with 48 noncases, and

concordant with 50 + 100 = 150 noncases.

When $X_1 = 0$ and $X_2 = 1$:

20 cases have lower $\hat{\text{P}}(\mathbf{X})$ than
2 + 48 = 50 noncases
i.e., 20 × 50 = **1,000 discordant** pairs

20 cases and 50 noncases have same $\hat{\text{P}}(\mathbf{X})$,
i.e., 20 × 50 = **1,000 tied pairs**

20 cases have higher $\hat{\text{P}}(\mathbf{X})$ than 100 noncases
i.e., 20 × 100 = **2,000 concordant** pairs

The 20 cases with $X_1 = 0$ and $X_2 = 1$ are

discordant with 2 + 48 = 50 noncases,

tied with 50 noncases, and

concordant with 100 noncases.

EXAMPLE (continued)

When $X_1 = 0$ and $X_2 = 0$:

$$\begin{cases} \text{20 cases have lower } \hat{P}(\mathbf{X}) \text{ than} \\ \quad 2 + 48 + 50 = 100 \text{ noncases} \\ \quad \text{i.e., } 20 \times 100 = \mathbf{2{,}000\ discordant} \text{ pairs} \\ \\ \text{20 cases and 100 noncases have same} \\ \hat{P}(\mathbf{X}), \text{ i.e., } 20 \times 100 = \mathbf{2{,}000\ tied\ pairs} \end{cases}$$

Finally, the 20 cases that did not have either risk factor $X_1 = 0$ and $X_2 = 0$ are

discordant with 2 + 48 + 50 = 100 noncases and

tied with 100 noncases.

total no. of concordant pairs:
$\mathbf{w} = 1{,}980 + 7{,}500 + 2{,}000 = \mathbf{11{,}480}$

total no. of tied pairs:
$\mathbf{z} = 20 + 2{,}400 + 1{,}000 + 2{,}000 = \mathbf{5{,}420}$

We now sum up all the above concordant and tied pairs, respectively, to obtain

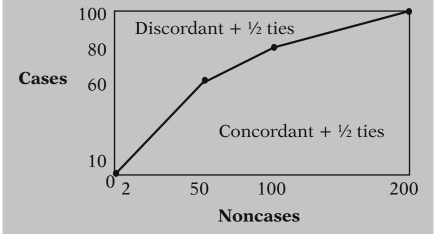
$\mathbf{w} = 11{,}480$ total concordant pairs and

$\mathbf{z} = 5{,}420$ total tied pairs.

$$\begin{aligned} \text{AUC} &= \frac{w + 0.5z}{n_\text{p}} \\ &= \frac{11{,}480 + 0.5(5{,}420)}{20{,}000} = \frac{14{,}190}{20{,}000} \\ &= \mathbf{0.7095} \end{aligned}$$

We then use w, z, and n_p to calculate the area under the ROC curve using the AUC formula, as shown on the left.

Geometrical Approach for Calculating AUC



To describe how to obtain this result geometrically, we first point out that with 100 cases and 200 noncases, the total number of case/noncase pairs (i.e., 100 × 200) can be geometrically represented by the rectangular area with height 100 and width 200 shown at the left.

ROC curve: scaled-up
(from 100% × 100% axes to 100 × 200 axes)
Y-axis: no. of cases testing +
X-axis: no. of noncases testing +

A scaled-up version of the ROC curve is superimposed within this area. Also, the values listed on the Y-axis (i.e., for cases) correspond to the number of cases testing positive at the cut-points used to plot the ROC curve. Similarly, the values listed on the X-axis (i.e., for noncases) correspond to the number of noncases testing positive at these same cut-points.

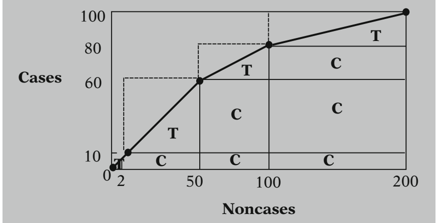
EXAMPLE (continued)

Concordant pairs: within area under ROC

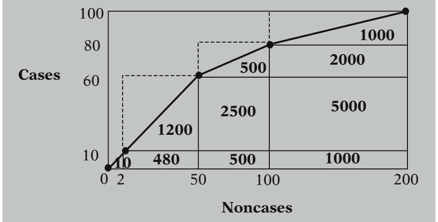
Discordant pairs: within area over ROC

Tied pairs, split equally over and under ROC

Within the above rectangle, the *concordant pairs* are represented by the area *under* the ROC curve while the *discordant pairs* are represented by the area *over* the ROC curve. The *tied pairs* and are split equally over and under the ROC curve (using the trapezoid rule).



To compute the actual area within the rectangle under the ROC curve, we can partition this area using sub-areas of rectangles and triangles as shown at the left. The areas denoted by **C** represent concordant pairs. The triangular areas denoted by **T** represent ½ of tied pairs.



Using the grids provided for the Y- and X-axes, the actual areas can be calculated as shown at the left. Note that an area labeled as **T** is calculated as ½ the corresponding rectangular area above and below the hypotenuse of a triangle that connects two consecutive cut points.

Sum of subareas under ROC

$$= 10 + 480 + 500 + 1{,}000 + 1{,}200 + \cdots + 1{,}000$$

$$= 14{,}190$$

Proportion of total rectangular area under ROC

$$= 14{,}190/20{,}000 = 0.7095 \ (= \text{AUC})$$

The sum of all the subareas under the curve is 14,190, whereas the total area in the rectangle of width 200 and height 100 is 200×100, or 20,000 (n_{p}). Therefore, the proportion of the total area taken up by the area under the ROC curve is 14,190 divided by 20,000 or 0.7095, which is the value calculated using the AUC formula.

EXAMPLE (continued)

Alternative calculation of sub-area: Rewrite

$$n_p = 100 \times 200$$

as

$$(10 + 50 + 20 + 20) \times (2 + 48 + 50 + 100)$$

Classification information for different cutpoints (c_p)

X_1	X_2	$\hat{P}(X)$	Cases	Noncases
-	-	c_0=1	0	0
1	1	c_1	10	2
1	0	c_2	50	48
0	1	c_3	20	50
0	0	c_4	20	100

An alternative way to obtain the sub-area values without having to geometrically calculate the each subarea can be obtained by rewriting the product formula for the total case/noncase pairs as shown at the left.

Each term in the sum on the left side of this product gives the number of cases with the same predicted risk (i.e., $\hat{P}(\mathbf{X})$) at one of the cut-points used to form the ROC. Similarly each term in the sum on the right side gives the number of noncases with the same $\hat{P}(\mathbf{X})$ at each cut-point.

$$\begin{aligned}
&(10 + 50 + 20 + 20) \times (2 + 48 + 50 + 100)\\
&= 20_t + 480_c + 500_c + 1{,}000_c\\
&\quad + 100_d + 2{,}400_t + 2{,}500_c + 5{,}000_c\\
&\quad + 40_d + 960_d + 1{,}000_t + 2{,}000_c\\
&\quad + 40_d + 960_d + 1{,}000_d + 2{,}000_t
\end{aligned}$$

We then multiply the two partitioned terms in the product formula to obtain 16 different terms, as shown at the left. Those terms identified with the subscript "t" denote tied pairs, those terms with the subscript "c" denote concordant pairs, and those terms with the subscript "d" denote discordant pairs.

Same values as in geometrical diagram

$480_c + 500_c + 1{,}000_c + 2{,}500_c + 5{,}000_c + 2{,}000_c$
$=$ **11,480 concordant pairs** ($=$**w**)

The six values with the subscript "c" are exactly the same as the six concordant areas shown in the geometrical diagram given earlier. The sum of these six values, therefore, gives the total area *under* the ROC curve for concordant pairs (i.e., **w**).

Twice each triangular area

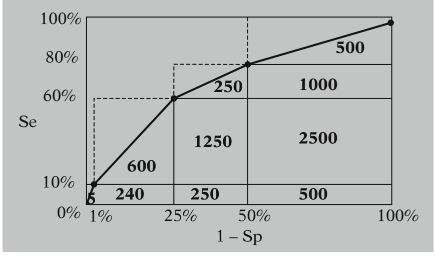
$20_t + 2{,}400_t + 1{,}000_t + 2{,}000_t$
$=$ **5,420 ties** ($=$**z**)

The four values with the subscript "t" are exactly twice the four triangular areas under the ROC curve. Their sum therefore gives twice the total tied pairs (i.e., **z**) *under* the ROC curve.

$100_d + 40_d + 960_d + 40_d + 960_d + 1{,}000_d$
$=$ **3,100 discordant pairs**

The remaining six terms identify portions of the area *above* the ROC curve corresponding to discordant pairs. These are not used to compute AUC.

Rescaled Area



Note that we can rescale the height and width of the rectangle to 100% × 100%, which will portray the dimensions of the rectangular area in (Se × 1 − Sp) percent mode. To do this, the value in each subarea under the curve needs to be halved, as shown at the left.

EXAMPLE (continued)

Combined area under rescaled ROC

$= 5 + 240 + 250 + 500 + 600 + \cdots + 500$
$= \mathbf{7{,}095}$

Proportion of total area under rescaled ROC

$= 7{,}095/10{,}000 = \mathbf{0.7095}$ $(= \text{AUC})$

The combined area under the rescaled ROC curve is then 7,095, which represents a proportion of 0.7095 of the total rectangular area of 10,000.

V. Example from Study on Screening for Knee Fracture

EXAMPLE

- Logistic model
- n = 348 ER patients
- Complaint: blunt knee trauma
- X-rayed for knee fracture

A logistic regression model was used in the analysis of a dataset containing information from 348 patients who entered an emergency room (ER) complaining of blunt knee trauma, and who subsequently were X-rayed for possible knee fracture (Tigges et al., 1999).

- Study purpose: use covariates to screen for decision to perform X-ray

The purpose of the analysis was to assess whether a patient's pattern of covariates could be used as a screening test before performing the X-ray.

- 1.3 million people per year visit ER with blunt knee trauma
- Substantial total cost for X-rays

Since 1.3 million people visit North American ER departments annually complaining of blunt knee trauma, the total cost associated with even a relatively inexpensive test such as a knee X-ray (about $200 for each X-ray) may be substantial.

Outcome variable:
FRACTURE = knee fracture status
(1 = yes, 0 = no)
Predictor variables:
FLEX = ability to flex knee
(0 = yes, 1 = no)
WEIGHT = ability to put weight on knee (0 = yes, 1 = no)
AGECAT = patient's age
($0 = \text{age} < 55$,
$1 = \text{age} \geq 55$)
HEAD = injury to knee head
(0 = no, 1 = yes)
PATELLAR = injury to patella
(0 = no, 1 = yes)

The variables considered in this analysis are listed at the left. The outcome variable is called FRACTURE, which represents a binary variable for knee fracture status.

The five predictor variables are FLEX, WEIGHT, AGECAT, HEAD, and PATELLAR, and are defined at the left.

EXAMPLE (continued)

Logistic Model:

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \beta_0 &+ \beta_1\text{FLEX}\\ &+ \beta_2\text{WEIGHT}\\ &+ \beta_3\text{AGECAT}\\ &+ \beta_4\text{HEAD}\\ &+ \beta_5\text{PATELLAR}\end{aligned}$$

The logistic model used in the analysis is shown at the left, and includes all five predictor variables. Although some of these predictors could have been evaluated for significance, we report here only on the ability of the 5-variable model to discriminate cases (fracture = 1) from noncases (fracture = 0).

Results shown below based on SAS's LOGISTIC procedure (but can also use STATA or SPSS)

We summarize the results of this analysis based on using SAS's LOGISTIC procedure, although the analysis could alternatively have been carried out using either STATA or SPSS (see Computer Appendix for computer code and output).

Fitted Logistic Regression Model:

```
                            Std      Wald      Pr >
Parameter  DF  Estimate     Err      ChiSq     ChiSq
Intercept   1   −3.4657   0.4118   70.8372    <.0001
FLEX        1    0.5277   0.3743    1.9877    0.1586
WEIGHT      1    1.5056   0.4093   13.5320    0.0002
AGECAT      1    0.5560   0.3994    1.9376    0.1639
HEAD        1    0.2183   0.3761    0.3367    0.5617
PATELLAR    1    0.6268   0.3518    3.1746    0.0748
```

The output showing the fitted model is now shown at the left.

- FLEX, WEIGHT, and HEAD have nonsignif Wald statistics.
- BW elimination would simplify model
- Focus for now on full model

Notice that three of the variables (FLEX, AGECAT, and HEAD) in the model have nonsignificant Wald tests, indicating backward elimination would result in removal of one or more of these variables, e.g., HEAD would be eliminated first, since it has the highest Wald *P*-value (0.5617). Nevertheless, we will focus on the full model for now, assuming that we wish to use all five predictors to carry out the screening.

Classification Table

Prob	Correct		Incorrect		Percentages			
Level	Event	Non-Event	Event	Non-Event	Correct	Se	Sp	1 – Sp †
0.000	45	0	303	0	12.9	**100.0**	0.0	**100.0**
0.050	39	93	210	6	37.9	**86.7**	30.7	**69.3**
0.100	36	184	119	9	63.2	**80.0**	60.7	**39.3**
0.150	31	200	103	14	66.4	**68.9**	66.0	**34.0**
0.200	32	235	68	23	73.9	**48.9**	77.6	**22.4**
0.250	16	266	37	29	81.0	**35.6**	87.8	**12.2**
0.300	6	271	32	39	79.6	**13.3**	89.4	**10.6**
0.350	3	297	6	42	86.2	**6.7**	98.0	**2.0**
0.400	3	301	2	42	87.4	**6.7**	99.3	**0.7**
0.450	2	301	2	43	87.1	**4.4**	99.3	**0.7**
0.500	0	303	0	45	87.1	**0.0**	100.0	**0.0**
0.550	0	303	0	45	87.1	**0.0**	100.0	**0.0**
0.600	0	303	0	45	87.1	**0.0**	100.0	**0.0**
0.650	0	303	0	45	87.1	**0.0**	100.0	**0.0**
0.700	0	303	0	45	87.1	**0.0**	100.0	**0.0**
0.750	0	030	0	45	87.1	**0.0**	100.0	**0.0**
0.800	0	030	0	45	87.1	**0.0**	100.0	**0.0**
0.850	0	303	0	45	87.1	**0.0**	100.0	**0.0**
0.900	0	303	0	45	87.1	**0.0**	100.0	**0.0**
0.950	0	303	0	45	87.1	**0.0**	100.0	**0.0**
1.000	0	303	0	45	87.1	**0.0**	100.0	**0.0**

† 1 – Sp is not automatically output in SAS; s LOGISTIC

We now show the classification table that uses the patients' predicted outcome probabilities obtained from the fitted logistic model to screen each patient. The probability levels (first column) are prespecified cut points (in increments of 0.05) requested in the model statement.

For example, in the third row, the cut-point is 0.100. If this cut-point is used for screening, then any patient whose predicted probability is greater than 0.100 will test positive for knee fracture on the screening test and therefore will receive an X-ray.

EXAMPLE (continued)

$c_p = 0.100$:

$\text{Se} = 36/45 = 0.80$

$\text{Sp} = 184/303 = 0.607$

$1 - \text{Sp} = 0.393$

Notice that at cut-point 0.100, 36 of 45 true events were correctly classified as events, and 9 of 45 were incorrectly classified as nonevents; also 184 of 303 true nonevents were correctly classified as nonevents, and 119 of 303 were incorrectly classified as events. The sensitivity (Se) for this row is 36/45, or 80%, the specificity (Sp) is 184/303, or 60.7%, so that 1 − Sp is 39.3%. Thus, in this row (cut-pt 0.100), the Se is larger than 1−Sp, which indicates good discrimination (for this cut point).

$\text{Se} = 0.80 > 1 - \text{Sp} = 0.393$
(good discrimination)

$\text{Se} \geq 1 - \text{Sp}$ for all cut-points,
where
$\text{Se} = 1 - \text{Sp} = 0$ for $c_p \geq 0.500$

We can also see from this table that Se is at least as large as 1 − Sp for all cut-pts. Notice, further, that once the cut-pt reaches 0.5 (and higher), none of the 45 true cases are correctly classified as cases (Se = 0) whereas all 303 true noncases are correctly classified as noncases (Sp = 1 and 1 − Sp = 0).

**Edited Output (SAS ProcLogistic)-
Association of Predicted Probabilities and Observed Responses**

Percent Concordant	71.8	Somers' D	0.489
Percent Discordant	22.9	Gamma	0.517
Percent Tied	5.3	Tau-a	0.111
Pairs	13635	**c**	0.745

Additional output obtained from SAS's Logistic procedure is shown at the left. This output contains information and statistical measures related to the ROC curve for the fitted model.

c = AUC
Somer's D, Gamma, and Tau-a:
other measures of discrimination

The "**c**" statistic of 0.745 in this output gives the area under the ROC curve, i.e., AUC, that we described earlier. The Somers' D, Gamma, and Tau-a are other measures of discrimination computed for the fitted model.

c, Somer's D, Gamma, and Tau-a:

(ranked) correlations between
observed outcomes ($Y_i = 0$ or 1)
and
predicted probabilities ($\hat{P}(\mathbf{X}_i)$)

Each of these measures involves different ways to compute a correlation between ranked (i.e., ordered) observed outcomes ($Y_i = 0$ or 1) and ranked predicted probabilities ($\hat{P}(\mathbf{X}_i)$). A high correlation indicates that higher predicted probabilities obtained from fitting the model correspond to true cases ($Y_i = 1$) whereas lower predicted probabilities correspond to true noncases ($Y_i = 0$), hence good discrimination.

Percent Concordant: 100 w/n_p
Percent Discordant: 100 d/n_p
Percent Tied: 100 z/n_p
Pairs: $n_p = n_1 \times n_0$,

The formulae for each measure are derived from the information provided on the left side of the above output. The definitions of each of the latter items are shown at the left. Note that w, z, and n_p were defined in the previous section for the formula for the AUC (i.e., **c**).

where
w = no. of case/noncase pairs for which

$$\hat{\mathrm{P}}(\mathbf{X}_{\mathbf{case}}) > \hat{\mathrm{P}}(\mathbf{X}_{\mathbf{noncase}})$$

d = no. of case/noncase pairs for which

$$\hat{\mathrm{P}}(\mathbf{X}_{\mathbf{noncase}}) > \hat{\mathrm{P}}(\mathbf{X}_{\mathbf{case}})$$

z = no. of case/noncase pairs for which

$$\hat{\mathrm{P}}(\mathbf{X}_{\mathbf{case}}) = \hat{\mathrm{P}}(\mathbf{X}_{\mathbf{noncase}})$$

Formulae for discrimination measures:

$$\boldsymbol{c} = \frac{w + 0.5z}{n_{\mathrm{p}}} = \mathrm{AUC}$$

$$\text{Somer's D} = \frac{w - d}{n_{\mathrm{p}}}$$

$$\text{Gamma} = \frac{w - d}{w + d}$$

$$\text{Tau-a} = \frac{w - d}{0.5\sum_i Y_i(\sum_i Y_i - 1)}$$

Using the notation just described, the formulae for these discrimination measures are shown at the left, with the first of these formulae (for the AUC) provided in the previous section.

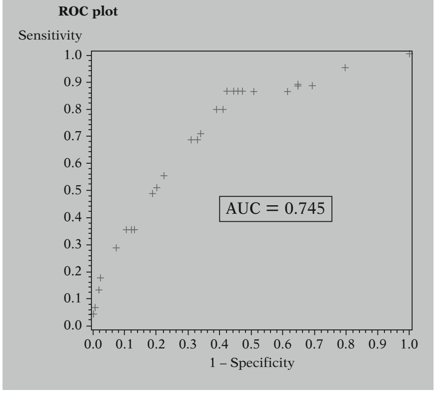
EXAMPLE

$$c = \frac{w + 0.5z}{n_{\mathrm{p}}}$$
$$= \frac{13{,}635(.718) + 0.5(13{,}635)(.053)}{13{,}635}$$
$$= 0.745$$

AUC = 0.745 ⇒ Fair discrimination (grade C)

The calculation of the AUC for the fitted model is shown at the left. The value for w in this formula is 13,635(.718), or 9,789.91 and the value for z is (13,635)(.053), or 722.655.

Based on the AUC result of 0.745 for these data, there is evidence of fair (Grade C) discrimination using the fitted model.



A plot of the ROC curve for these data can also be obtained and is shown here. Notice that the points on the plot that represent the coordinates of Se by 1 − Sp at different cut-pts have not been connected by the program. Nevertheless, it is possible to fit a cubic regression to the plotted points of sensitivity by 1 − specificity (not shown, but see Computer Appendix).

EXAMPLE (continued)

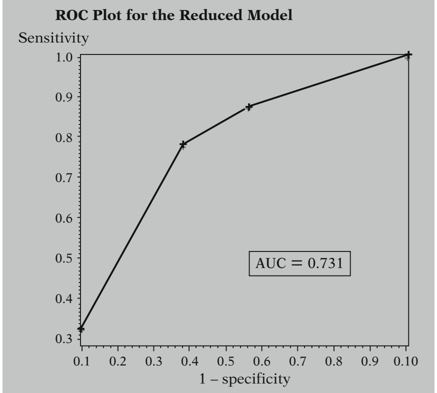
Backward elimination:
Step 1: Drop HEAD
(highest *P*-value 0.5617)
Step 2: Drop AGECAT
(highest *P*-value 0.2219)
Step 3: Drop FLEX
(highest *P*-value 0.1207)
Step 4: Keep WEIGHT or PATELLAR
(highest *P*-value 0.0563)

Reduced Model After BW Elimination

Parameter	DF	Estimate	Std Err	Wald ChiSq	Pr > ChiSq
Intercept	1	−3.1790	0.3553	80.0692	<.0001
WEIGHT	1	1.7743	0.3781	22.0214	<.0001
PATELLAR	1	0.6504	0.3407	3.6437	0.0563

Association of Predicted Probabilities and Observed Responses

Percent Concordant	61.4	Somers' D	0.463
Percent Discordant	15.2	Gamma	0.604
Percent Tied	23.4	Tau-a	0.105
Pairs	14,214	*c*	0.731



Reduced model (2 *X*s)		Full model (5 *X*s)
$2^2 = 4$ covariate patterns		$2^5 = 32$ covariate patterns
4 cut-pts	≤	28 cut-pts
$\text{AUC}_{\text{Reduced}} = 0.731$	≤	$\text{AUC}_{\text{Full}} = 0.745$

In general:
Model 1 is **nested** within Model 2

$$\Downarrow$$

$$\text{AUC}_{\text{Model 1}} \leq \text{AUC}_{\text{Model 2}}$$

Recall that the previously shown output gave Wald statistics for HEAD, AGECAT, and FLEX that were nonsignificant. A backward elimination approach that begins by dropping the least significant of these variables (i.e., HEAD), refitting the model, and dropping additional nonsignificant variables results in a model that contains only two predictor variables, WEIGHT and PATELLAR. (Note that we are treating all predictor variables as exposure variables.)

The fitted logistic model that involves only WEIGHT and PATELLAR is shown at the left.

We also show the discrimination measures that result for this model, including the *c* (= AUC) statistic. The *c* statistic here is 0.731, which is slightly smaller than the *c* statistic of 0.745 obtained for the full model. The reduced model has slightly less discriminatory power than the full model. (See Hanley (1983) for a statistical test of significance between two or more AUCs.)

The ROC plot for the reduced model is shown here.

Notice that there are fewer cut-pts plotted on this graph than on the ROC plot for the full model (previous page). The reason is that the number of possible cut-pts for a given model is always equal to or less than the number of covariate patterns (i.e., distinct combinations of predictors) defined by the model.

The reduced model (with only two binary predictors) contains four (=2^2) covariate patterns whereas the full model (with 5 binary predictors) contains 32 (=2^5) covariate patterns.

Moreover, because the reduced model is nested within the full model, the AUC for the reduced model will always be smaller than the AUC for the full model, i.e., similar to the characteristics of R^2 in linear regression. That's the case here, since the AUC is 0.731 for the reduced model compared to 0.745 for the full model.

EXAMPLE (continued)

Model 3: HEAD, AGECAT, and FLEX

$$AUC_{\text{Model 3}} = 0.660$$

Reduced Model: WEIGHT and PATELLAR

$$AUC_{\text{Model 2}} = 0.731$$

Thus, Model 2 (fewer variables) discriminates better than Model 3 (more variables)

Note, however, if two models that are not nested are compared, there is no guarantee which model will have a larger AUC. For example, the model that contains only the three variables that were dropped, namely HEAD, AGECAT, and FLEX has an AUC of 0.660, which is smaller than the AUC of 0.731 obtained for the two variable (reduced) model involving WEIGHT and PATELLAR.

VI. SUMMARY

DP = discriminatory performance of a binary logistic model

This presentation is now complete. We have described how to assess *discriminatory performance (DP)* of a binary logistic model.

Good DP: model discriminates **cases** ($Y = 1$) from **noncases** ($Y = 0$)

A model provides *good DP* if the covariates in the model help to predict (i.e., discriminate) which subjects will develop the outcome ($Y = 1$, or the *cases*) and which will not develop the outcome ($Y = 0$, or the *noncases*).

One approach:

Classification/Diagnostic Table

True (Observed) Outcome

	$\mathbf{c_p}$	$Y = 1$	$Y = 0$
Predicted	$Y = 1$	n_{TP}	n_{FP}
Outcome	$Y = 0$	n_{FN}	n_{TN}
		n_1	n_0

$\mathbf{c_p}$ = cut-point for classifying cases vs. noncases

Se = Pr(true+ | true C) = n_{TP}/n_1

Sp = Pr(true− | true NC) = n_{TN}/n_0

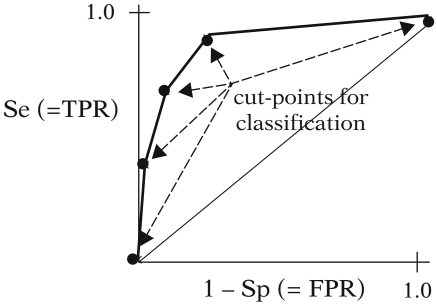
One way to measure **DP** is to consider the *sensitivity* (*Se*) *and specificity* (*Sp*) *from a classification table* that combines observed and predicted outcomes over all subjects. The closer both the sensitivity and specificity are to 1, the better is the discrimination.

Another approach:

Plot and/or **summary measure** based on a range of cut-points

An alternative way to measure *DP* involves a plot and/or summary measure based on a range of cut-points chosen for a given model.

ROC curve



A widely used plot is the *ROC curve,* which graphs the sensitivity by 1 minus the specificity for a range of cut-points. Equivalently, the ROC is a plot of the *true positive rate* (*TPR* = Se) *by* the *false positive rate* (*FPR* = 1 − Sp).

AUC = area under ROC curve

AUC = 1.0 ⇒ perfect **DP**

(Se = 1 − Sp)

AUC = 0.5 ⇒ no **DP**

A popular summary measure based on the ROC plot is the area under the ROC curve, or *AUC*. The larger the AUC, the better is the DP. An AUC of 1 indicates perfect DP and an AUC of 0.5 indicates no DP.

We suggest that you review the material covered in this chapter by reading the detailed outline that follows. Then do the practice exercises and test.

Up to this point, we have considered binary outcomes only. In the next two chapters, the standard logistic model is extended to handle outcomes with three or more categories.

Detailed Outline

I. Overview (pages 348–350)

A. Focus: how to assess discriminatory performance (*DP*) of a binary logistic model.

B. Considers how well the covariates in a given model help to predict (i.e., discriminate) which subjects will develop the outcome (Y = 1, or the *cases*) and which will not develop the outcome (Y = 0, or the *noncases*).

C. One way to measure *DP*: consider the *sensitivity (Se) and specificity (Sp) from a classification table* that combines true and predicted outcomes over all subjects.

D. An alternative way to measure *DP*: involves a plot (i.e., ROC curve) and/or summary measure (AUC) based on a range of cut-points chosen for a given model.

II. Assessing Discriminatory Performance using Sensitivity and Specificity Parameters (pages 350–354)

A. Classification Table

i. One way to assess DP.

ii. Combines true and predicted outcomes over all subjects.

iii. Cut-point (c_p) can be used with $\hat{P}(\mathbf{X})$ to predict whether subject is case or noncase:

- If $\hat{P}(\mathbf{X}) > c_p$, then predict subj **X** to be case; otherwise, predict subj **X** to be noncase.

B. Sensitivity (Se) and specificity (Sp)

i. Computed from classification table for fixed cut point.

ii. Se = proportion of truly diagnosed cases
$= \Pr(\text{true positive} \mid \text{true case}) = n_{TP}/n_1$

iii. Sp = proportion of falsely diagnosed
$= \Pr(\text{true negative} \mid \text{true noncase}) = n_{TN}/n_0$

iv. The closer both Se and Sp are to 1, the better is the discrimination.

v. Sp and Se values vary with c_p:

- c_p decreases from 1 to 0 ⇒ Se increases from 0 to 1, and Sp decreases from 1 to 0.
- Sp may change at a different rate than the Se depending on the model considered.

vi. 1 − Sp more appealing than Sp:

- Se and 1 − Sp both focus on predicted cases
- If good discrimination, would expect Se > 1 − Sp for all c_p

C. Pick a case and a noncase at random: what is probability that $\hat{P}(\mathbf{X_{case}}) > \hat{P}(\mathbf{X_{noncase}})$**?**

i. One approach: "collectively" determine whether Se exceeds 1 − Sp over several cut-points ranging between 0 and 1.

ii. Drawback: Se and Sp values are "summary statistics" over several subjects.

iii. Instead: use proportion of case, noncase pairs for which $\hat{P}(\mathbf{X_{case}}) \geq \hat{P}(\mathbf{X_{noncase}})$.

III. Receiver Operating Characteristic (ROC) Curves (pages 354–358)

A. ROC plots *sensitivity* ***(Se) by 1 − specifity (1 – Sp)*** values over all cut points.

i. Equivalently, ROC plots *true positive rate* (*TPR*) for cases *by* the *false positive rate* (*FPR*) for noncases.

B. ROC measures how well model predicts who will or will not have the outcome.

C. ROC provides numerical answer to question: for randomly case/noncase pair, what is probability that $\hat{P}(\mathbf{X_{case}}) \geq \hat{P}(\mathbf{X_{noncase}})$**?**

i. The answer: AUC = area under the ROC.

ii. The larger the area, the better is the discrimination.

iii. Two extremes:

AUC = 1 ⇒ perfect discrimination

AUC = 0.5 ⇒ no discrimination

D. Grading guidelines for AUC values:

0.90 − 1.0 = excellent discrimination (A); rarely observed

0.80 − 0.90 = good discrimination (B)

0.70 − 0.80 = fair discrimination (C)

0.60 − 0.70 = poor discrimination (D)

0.50 − 0.60 = failed discrimination (F)

E. Complete separation of points (CSP)

i. Occurs if all exposed subjects are cases and almost all unexposed subjects are noncases.

ii. CSP often found when AUC ≥ 0.90.

iii. CSP ⇒ impossible as well as unnecessary to fit a logistic model to the data.

IV. Computing the Area Under the ROC (pages 358–365)

A. General formula:

$$\text{AUC} = \frac{\text{no. of case/noncase pairs in which } \hat{\text{P}}(\mathbf{X_{case}}) \geq \hat{\text{P}}(\mathbf{X_{noncase}})}{\text{total \# case–control pairs}}$$

B. Calculation formula:

$$c(= \text{AUC}) = \frac{w + 0.5z}{n_\text{p}},$$

where

w = no. of case/noncase pairs for which $\hat{\text{P}}(\mathbf{X_{case}}) > \hat{\text{P}}(\mathbf{X_{noncase}})$

z = no. of case/noncase pairs for which $\hat{\text{P}}(\mathbf{X_{case}}) = \hat{\text{P}}(\mathbf{X_{noncase}})$

n_p = total no. of case/noncase pairs

C. Example of AUC calculation:

i. $n = 300$ subjects, $n_1 = 100$ true cases, $n_0 = 200$ true noncases

ii. Fit logistic model P(**X**) and compute $\hat{\text{P}}(\mathbf{X_i})$ for $i = 1, \ldots, 300$

iii. $n_\text{p} = n_1 \times n_0 = 100 \times 200 = 20{,}000$

iv. Suppose $w = 11{,}480$ (i.e., 57.4% of 20,000) and $z = 5{,}420$ (i.e., 27.1% of 20,000)

v. $c\,(= \text{AUC}) = \frac{w + 0.5z}{n_\text{p}} = \frac{11{,}480 + 0.5(5{,}420)}{20{,}000} = 0.7095,$ grade C (fair) discrimination

V. Examples from Study on Screening for Knee Fracture (pages 365–370)

A. Scenario: n = 348 ER patients; complaint: blunt knee trauma; X-rayed for knee fracture.

B. Study purpose: use covariates to screen for decision to perform X-ray.

C. Outcome variable: FRACTURE = knee fracture status (1 = yes, 0 = no)

Predictor variables:

FLEX = ability to flex knee (0 = yes, 1 = no)

WEIGHT = ability to put weight on knee (0 = yes, 1 = no)

AGECAT = patients age (0 = age < 55, 1 = age ≥ 55)

HEAD = injury to knee head (0 = no, 1 = yes)

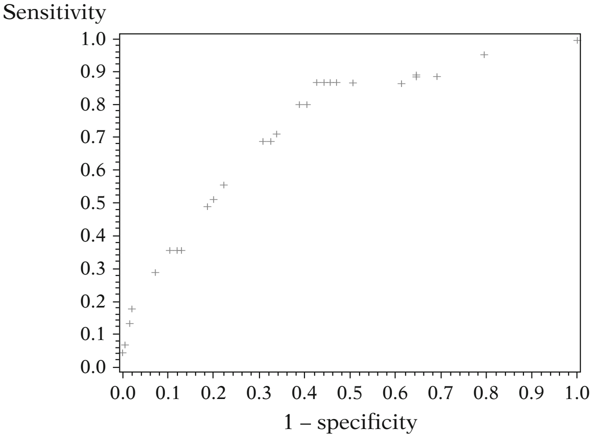
PATELLAR = injury to patella (0 = no, 1 = yes)

D. Logistic Model:

$$\text{logit } P(\mathbf{X}) = \beta_0 + \beta_1\text{FLEX} + \beta_2\text{WEIGHT} + \beta_3\text{AGECAT} + \beta_4\text{HEAD} + \beta_5\text{PATELLAR}$$

E. Results based on SAS's LOGISTIC procedure (but can also use STATA or SPSS).

F. ROC plot



G. AUC = 0.745 ⇒ Fair discrimination (Grade C)

H. Reduced Model

i. Why? Some nonsignificant regression coefficients in the full model

ii. Use backward elimination to obtain following reduced model:

$$\text{logit } P(\mathbf{X}) = \beta_0 + \beta_2\text{WEIGHT} + \beta_5\text{PATELLAR}$$

iii. AUC (Reduced model) $= 0.731 \leq$ AUC (Full model) $= 0.745$

iv. In general, for nested models, AUC(smaller model) $\leq$ AUC (larger model),

v. However, if models not nested, it is possible that AUC(model with fewer variables) $>$ AUC (model with more variables).

VI. Summary (page 371)

Practice Exercises

The following questions and computer information consider the Evans Country dataset on 609 white males that has been previously discussed and illustrated in earlier chapters of this text. Recall that the outcome variable is CHD status (1 = case, 0 = noncase), the exposure variable of interest is CAT status (1 = high CAT, 0 = low CAT), and the five control variables considered are AGE (continuous), CHL (continuous), ECG (0,1), SMK (0,1), and HPT (0,1).

The SAS output provided below was obtained for the following logistic model:

$$\begin{aligned}\text{Logit P}(\mathbf{X}) = \alpha &+ \beta_1\text{CAT} + \gamma_1\text{AGE} + \gamma_2\text{CHL} + \gamma_3\text{ECG}\\ &+ \gamma_4\text{SMK} + \gamma_5\text{HPT} + \delta_1\text{CC} + \delta_2\text{CH},\end{aligned}$$

where CC = CAT × CHL and CH = CAT × HPT

```
        Analysis of Maximum Likelihood Estimates
                           Standard    Wald
Parameter  DF   Estimate   Error       Chi-     Pr >
                                       Square   ChiSq
Intercept  1    −4.0497    1.2550     10.4125   0.0013
cat        1   −12.6894    3.1047     16.7055  <.0001
age        1     0.0350    0.0161      4.6936   0.0303
chl        1    −0.00545   0.00418     1.7000   0.1923
ecg        1     0.3671    0.3278      1.2543   0.2627
smk        1     0.7732    0.3273      5.5821   0.0181
hpt        1     1.0466    0.3316      9.9605   0.0016
cc         1     0.0692    0.0144     23.2020  <.0001
ch         1    −2.3318    0.7427      9.8579   0.0017
```

```
Association of Predicted Probabilities and Observed Responses
     Percent Concordant    78.6  Somers' D  0.578
     Percent Discordant    20.9  Gamma      0.580
     Percent Tied           0.5  Tau-a      0.119
     Pairs               38,198  c          0.789
```

```
                            Classification Table

               Correct         Incorrect                  Percentages
Prob                Non-             Non-              Sensi-  Speci-  False   False
Level      Event   event    Event   event   Correct   tivity  ficity   POS     NEG

0.000        71       0      538       0      11.7    100.0     0.0   88.3     .
0.020        68      35      503       3      16.9     95.8     6.5   88.1    7.9
0.040        67     127      411       4      31.9     94.4    23.6   86.0    3.1
0.060        60     226      312      11      47.0     84.5    42.0   83.9    4.6
0.080        54     326      212      17      62.4     76.1    60.6   79.7    5.0
0.100        50     393      145      21      72.7     70.4    73.0   74.4    5.1
0.120        41     425      113      30      76.5     57.7    79.0   73.4    6.6
0.140        37     445       93      34      79.1     52.1    82.7   71.5    7.1
0.160        34     463       75      37      81.6     47.9    86.1   68.8    7.4
0.180        34     477       61      37      83.9     47.9    88.7   64.2    7.2
0.200        31     495       43      40      86.4     43.7    92.0   58.1    7.5
0.220        29     504       34      42      87.5     40.8    93.7   54.0    7.7
0.240        28     509       29      43      88.2     39.4    94.6   50.9    7.8
0.260        27     514       24      44      88.8     38.0    95.5   47.1    7.9
0.280        25     519       19      46      89.3     35.2    96.5   43.2    8.1
0.300        23     525       13      48      90.0     32.4    97.6   36.1    8.4
```

```
                  Classification Table (continued)
            Correct       Incorrect            Percentages
Prob             Non-          Non-           Sensi- Speci-  False    False
Level    Event  event  Event  event  Correct tivity  ficity   POS     NEG

0.320     23     526    12     48     90.1    32.4    97.8   34.3     8.4
0.340     22     528    10     49     90.3    31.0    98.1   31.3     8.5
0.360     21     529     9     50     90.3    29.6    98.3   30.0     8.6
0.380     21     529     9     50     90.3    29.6    98.3   30.0     8.6
0.400     18     529     9     53     89.8    25.4    98.3   33.3     9.1
0.420     18     531     7     53     90.1    25.4    98.7   28.0     9.1
0.440     18     531     7     53     90.1    25.4    98.7   28.0     9.1
0.460     18     531     7     53     90.1    25.4    98.7   28.0     9.1
0.480     18     531     7     53     90.1    25.4    98.7   28.0     9.1
0.500     18     532     6     53     90.3    25.4    98.9   25.0     9.1
0.520     18     532     6     53     90.3    25.4    98.9   25.0     9.1
0.540     16     532     6     55     90.0    22.5    98.9   27.3     9.4
0.560     16     532     6     55     90.0    22.5    98.9   27.3     9.4
0.580     15     532     6     56     89.8    21.1    98.9   28.6     9.5
0.600     13     533     5     58     89.7    18.3    99.1   27.8     9.8
0.620     11     534     4     60     89.5    15.5    99.3   26.7    10.1
0.640     10     535     3     61     89.5    14.1    99.4   23.1    10.2
0.660     10     535     3     61     89.5    14.1    99.4   23.1    10.2
0.680     10     535     3     61     89.5    14.1    99.4   23.1    10.2
0.700     10     536     2     61     89.7    14.1    99.6   16.7    10.2
0.720      9     536     2     62     89.5    12.7    99.6   18.2    10.4
0.740      8     536     2     63     89.3    11.3    99.6   20.0    10.5
0.760      8     536     2     63     89.3    11.3    99.6   20.0    10.5
0.780      8     536     2     63     89.3    11.3    99.6   20.0    10.5
0.800      8     536     2     63     89.3    11.3    99.6   20.0    10.5
0.820      6     536     2     65     89.0     8.5    99.6   25.0    10.8
0.840      6     537     1     65     89.2     8.5    99.8   14.3    10.8
0.860      5     537     1     66     89.0     7.0    99.8   16.7    10.9
0.880      5     537     1     66     89.0     7.0    99.8   16.7    10.9
0.900      5     537     1     66     89.0     7.0    99.8   16.7    10.9
0.920      5     537     1     66     89.0     7.0    99.8   16.7    10.9
0.940      4     538     0     67     89.0     5.6   100.0    0.0    11.1
0.960      3     538     0     68     88.8     4.2   100.0    0.0    11.2
0.980      3     538     0     68     88.8     4.2   100.0    0.0    11.2
1.000      0     538     0     71     88.3     0.0   100.0     .     11.7
```

1. Using the above output:
   a. Give a formula for calculating the estimated probability $\hat{P}(\mathbf{X}^*)$ of being a case (i.e., CHD = 1) for a subject ($\mathbf{X}^*$) with the following covariate values: CAT = 1, AGE = 50, CHL = 200, ECG = 0, SMK = 0, HPT = 0?

      [Hint: $\hat{P}(\mathbf{X}^*) = 1/\{1 + \exp[-\text{logit } \hat{P}(\mathbf{X}^*)]\}$ where logit $\hat{P}(\mathbf{X}^*)$ is calculated using the estimated regression coefficients for the fitted model.]

   b. Compute the value of $\hat{P}(\mathbf{X}^*)$ using your answer to question 1a.

   c. If a discrimination cut-point of 0.200 is used to classify a subject as either a case or a noncase, how would you classify subject $\mathbf{X}^*$ based on your answer to question 1b.

   d. With a cut-point of 0.000, the sensitivity of the screening test is 1.0 (or 100% – see first row). Why does the sensitivity of a test have to be 100% if the cut point is 0? (assume there is at least one true event)

e. Notice for this data, as the cut-point gets larger the specificity also gets larger (or stays the same). For example, a cut-point of 0.200 yields a specificity of 92.0% while a cut-point of 0.300 yields a specificity of 97.6%. Is it possible (using different data) that an increase of a cut-point could actually decrease the specificity? Explain.

f. In the classification table provided above, a cut-point of 0.200 yields a false positive percentage of 58.1% whereas 1 minus the specificity at this cut point is 8.0%. Since 1 minus specificity percentage is defined as 100 times the proportion of true non-cases that are falsely classified as cases, i.e., the numerator in this proportion is the number of false-positive noncases, why is not the false positive percentage (58.1%) shown in the output equal to 1 minus specificity (8.0%)? Is the computer program in error?

2. Based on the output,

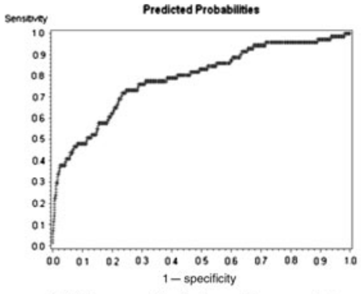
a. What is the area under the ROC curve? How would you grade this area in terms of the discriminatory power of the model being fitted?

b. In the output provided under the heading "Association of Predicted Probabilities and Observed Responses," the number of pairs is 38,198. How is this number computed?

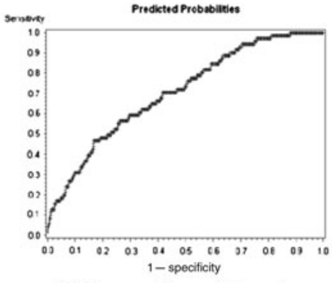
c. In the output provided under the same heading in question 2b, how are the Percent Concordant and the Percent Tied computed?

d. Using the information given by the number of pairs, the Percent Concordant and the Percent Tied described in parts (b) and (c), compute the area under the ROC curve (AUC) and verify that it is equal to your answer to part 2a.

e. The ROC curves for the interaction model described above and the no interaction model that does not contain the CC or CH (interaction) variables are shown below. The area under the ROC curve for the no-interaction model is 0.705. Why is the latter AUC less than the AUC for the interaction model?



ROC curve for interaction model



ROC curve for no interaction model

Test

The following questions and computer output consider a data from a cross-sectional study carried out at Grady Hospital in Atlanta, Georgia involving 289 adult patients seen in an emergency department whose blood cultures taken within 24 hours of admission were found to have Staph aureus infection (Rezende et al., 2002). Information was obtained on several variables, some of which were considered risk factors for methicillin resistance (MRSA). The outcome variable is MRSA status (1 = yes, 0 = no), and covariates of interest included the following variables: PREVHOSP (1 = previous hospitalization, 0 = no previous hospitalization), AGE (continuous), GENDER (1 = male, 0 = female), and PAMU (1 = antimicrobial drug use in the previous 3 months, 0 = no previous antimicrobial drug use).

The SAS output provided below was obtained for the following logistic model:

$$\text{Logit P}(\mathbf{X}) = \alpha + \beta_1\text{PREVHOSP} + \beta_2\text{AGE} + \beta_3\text{GENDER} + \beta_4\text{PAMU}$$

```
       Analysis of maximum likelihood estimates
                                         Wald
                              Standard   Chi-
Parameter  DF  Estimate  Error      Square    Pr > ChiSq

Intercept   1  -5.0583   0.7643     43.8059    <.0001
PREVHOSP    1   1.4855   0.4032     13.5745    0.0002
AGE         1   0.0353   0.00920    14.7004    0.0001
gender      1   0.9329   0.3418      7.4513    0.0063
pamu        1   1.7819   0.3707     23.1113    <.0001

                   Odds Ratio Estimates
               Point              95% Wald
     Effect    Estimate    Confidence    Limits

     PREVHOSP    4.417          2.004        9.734
     AGE         1.036          1.017        1.055
     gender      2.542          1.301        4.967
     pamu        5.941          2.873       12.285

    Association of Predicted Probabilities and Ob-
                    served Responses

  Percent Concordant     83.8    Somers' D    0.681
  Percent Discordant     15.8        Gamma    0.684
  Percent Tied            0.4        Tau-a    0.326
  Pairs                 19950            c    0.840
```

```
                         Classification Table
              Correct        Incorrect              Percentages
Prob               Non-             Non-            Sensi-  Speci-  False    False
Level     Event   event   Event   event   Correct  tivity  ficity   POS      NEG

0.000       114       0     175       0     39.4    100.0     0.0   60.6       .
0.020       114       2     173       0     40.1    100.0     1.1   60.3     0.0
0.040       113      19     156       1     45.7     99.1    10.9   58.0     5.0
0.060       110      38     137       4     51.2     96.5    21.7   55.5     9.5
0.080       108      58     117       6     57.4     94.7    33.1   52.0     9.4
0.100       107      73     102       7     62.3     93.9    41.7   48.8     8.8
0.120       107      75     100       7     63.0     93.9    42.9   48.3     8.5
0.140       106      81      94       8     64.7     93.0    46.3   47.0     9.0
0.160       106      89      86       8     67.5     93.0    50.9   44.8     8.2
0.180       106      91      84       8     68.2     93.0    52.0   44.2     8.1
0.200       106      93      82       8     68.9     93.0    53.1   43.6     7.9
0.220       102      99      76      12     69.6     89.5    56.6   42.7    10.8
0.240       101     100      75      13     69.6     88.6    57.1   42.6    11.5
0.260        99     105      70      15     70.6     86.8    60.0   41.4    12.5
0.280        98     106      69      16     70.6     86.0    60.6   41.3    13.1
0.300        98     107      68      16     70.9     86.0    61.1   41.0    13.0
0.320        96     115      60      18     73.0     84.2    65.7   38.5    13.5
0.340        96     117      58      18     73.7     84.2    66.9   37.7    13.3
0.360        95     119      56      19     74.0     83.3    68.0   37.1    13.8
0.380        92     120      55      22     73.4     80.7    68.6   37.4    15.5
0.400        91     121      54      23     73.4     79.8    69.1   37.2    16.0
0.420        91     125      50      23     74.7     79.8    71.4   35.5    15.5
0.440        91     127      48      23     75.4     79.8    72.6   34.5    15.3
0.460        89     129      46      25     75.4     78.1    73.7   34.1    16.2
0.480        85     134      41      29     75.8     74.6    76.6   32.5    17.8
0.500        82     138      37      32     76.1     71.9    78.9   31.1    18.8
0.520        81     140      35      33     76.5     71.1    80.0   30.2    19.1
0.540        79     141      34      35     76.1     69.3    80.6   30.1    19.9
0.560        75     145      30      39     76.1     65.8    82.9   28.6    21.2
0.580        73     147      28      41     76.1     64.0    84.0   27.7    21.8
0.600        70     151      24      44     76.5     61.4    86.3   25.5    22.6
0.620        66     153      22      48     75.8     57.9    87.4   25.0    23.9
0.640        61     156      19      53     75.1     53.5    89.1   23.8    25.4
0.660        55     160      15      59     74.4     48.2    91.4   21.4    26.9
0.680        48     163      12      66     73.0     42.1    93.1   20.0    28.8
0.700        42     165      10      72     71.6     36.8    94.3   19.2    30.4
0.720        34     167       8      80     69.6     29.8    95.4   19.0    32.4
0.740        32     171       4      82     70.2     28.1    97.7   11.1    32.4
0.760        29     171       4      85     69.2     25.4    97.7   12.1    33.2
0.780        25     171       4      89     67.8     21.9    97.7   13.8    34.2
0.800        17     172       3      97     65.4     14.9    98.3   15.0    36.1
0.820        12     173       2     102     64.0     10.5    98.9   14.3    37.1
0.840        11     174       1     103     64.0      9.6    99.4    8.3    37.2
0.860         6     174       1     108     62.3      5.3    99.4   14.3    38.3
0.880         5     174       1     109     61.9      4.4    99.4   16.7    38.5
0.900         0     175       0     114     60.6      0.0   100.0      .    39.4
0.920         0     175       0     114     60.6      0.0   100.0      .    39.4
0.940         0     175       0     114     60.6      0.0   100.0      .    39.4
0.960         0     175       0     114     60.6      0.0   100.0      .    39.4
0.980         0     175       0     114     60.6      0.0   100.0      .    39.4
1.000         0     175       0     114     60.6      0.0   100.0      .    39.4
```

Questions based on the above information begin on the next page.

1. For a discrimination cut-point of **0.300** in the Classification Table provided above,

   a. fill in the table below to show the cell frequencies for the number of true positives (n_{TP}), false positives (n_{FP}), true negatives (n_{TN}), and false negatives (n_{FN}):

		True (Observed) Outcome	
	$\boldsymbol{c_{\text{p}} = 0.30}$	$Y = 1$	$Y = 0$
Predicted	$Y = 1$	$n_{\text{TP}} =$	$n_{\text{FP}} =$
Outcome	$Y = 0$	$n_{\text{FN}} =$	$n_{\text{TN}} =$
		$n_1 = 114$	$n_0 = 175$

   b. Using the cell frequencies in the table of part 1a, compute *in percentages* the sensitivity, specificity, 1 − specificity, false positive, and false negative values, and verify that these results are identical to the results shown in the Classification Table for cut-point 0.300:

      Sensitivity % =
      Specificity % =
      1 − specificity % =
      False positive % =
      False negative % =

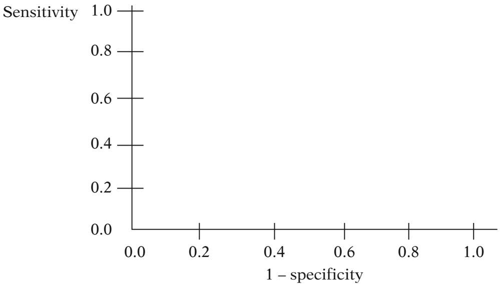
   c. Why are the 1 − specificity and false positive percentages *not* identical even though they both use the (same) number of false positive subjects in their calculation?

   d. How is the value of 70.9 in the column labeled "Correct" computed and how can this value be interpreted?

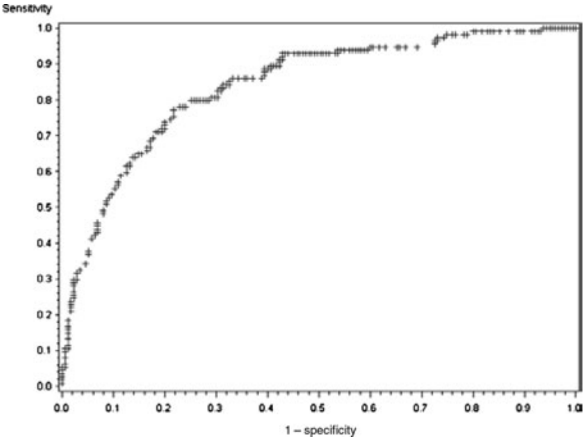
   e. How do you interpret values for sensitivity and specificity obtained for the cut-point of 0.300 in terms of how well the model discriminates cases from noncases?

   f. What is the drawback to (exclusively) using the results for the cut-point of 0.300 to determine how well the model discriminates cases from noncases?

2. Using the following graph, plot the points on the graph that would give the portion of the ROC curve that corresponds to the following cut-points: 0.000, 0.200, 0.400, 0.600, 0.800, and 1.000

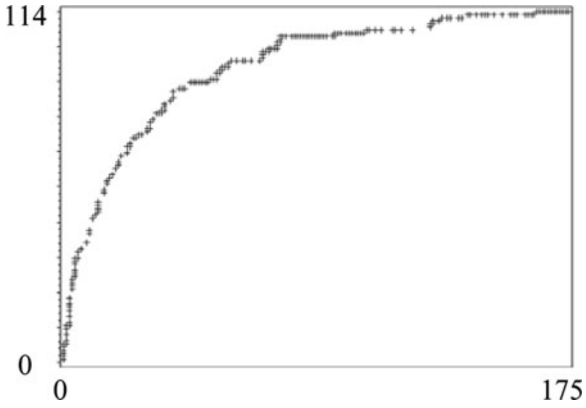


3. The ROC curve obtained for the model fitted to these data is shown below.



   a. Verify that the plots you produced to answer question 2 correspond to the appropriate points on the ROC curve shown here.
   b. Based on the output provided, what is the area under the ROC curve? How would you grade this area in terms of the discriminatory power of the model being fitted?
   c. In the output provided under the heading "Association of Predicted Probabilities and Observed Responses," the number of pairs is 19,950. How is this number computed?
   d. Using the information given by the number of pairs, the Percent Concordant, and the Percent Tied in the output under the heading "Association of Predicted Probabilities and Observed Responses," compute the area under the ROC curve (AUC), and verify that it is equal to your answer to part 3b.

4. Consider the following figure that superimposes the ROC curve within the rectangular area whose height is equal to the number of MRSA cases (114) and whose width is equal to the number of MRSA noncases (175).



a. What is the area within the entire rectangle and what does it have in common with the formula for the area under the ROC curve?

b. Using the AUC calculation formula, what is the area under the ROC curve superimposed on the above graph? How do you interpret this area?

5. Below is additional output providing goodness of fit information and the Hosmer-Lemeshow test for the model fitted to the MRSA dataset. The column labeled as "Group" lists the deciles of risk, ordered from smallest to largest, e.g., decile 10 contains 23 patients who had had the highest 10% of predicted probabilities.

```
                     mrsa = 1            mrsa = 0
Group  Total  Observed  Expected  Observed  Expected

  1      29       1        0.99      28       28.01
  2      31       5        1.95      26       29.05
  3      29       2        2.85      27       26.15
  4      29       5        5.73      24       23.27
  5      30      10        9.98      20       20.02
  6      31      12       14.93      19       16.07
  7      29      16       17.23      13       11.77
  8      29      20       19.42       9        9.58
  9      29      22       21.57       7        7.43
 10      23      21       19.36       2        3.64

             Hosmer and Lemeshow Goodness-of-
                         Fit Test

            Chi-Square      DF      Pr > ChiSq
            7.7793           8        0.4553
```

a. Based on the above output, does the model fit the data? Explain briefly.

b. What does the distribution of the number of observed cases and observed noncases over the 10 deciles indicate about how well the model discriminates cases from noncases? Does your answer coincide with your answer to question 3b in terms of the discriminatory power of the fitted model?

c. Suppose the distribution of observed and expected cases and noncases was given by the following table:

```
          Partition for the Hosmer and Lemeshow Test
                        mrsa = 1             mrsa = 0
Group  Total  Observed  Expected  Observed  Expected

 1      29       10        0.99      19       28.01
 2      31       11        1.95      20       29.05
 3      29       11        2.85      18       26.15
 4      29       11        5.73      18       23.27
 5      30       12        9.98      18       20.02
 6      31       12       14.93      19       16.07
 7      29       12       17.23      17       11.77
 8      29       13       19.42      16        9.58
 9      29       13       21.57      16        7.43
10      23        9       19.36      14        3.64
```

What does this information indicate about how well the model discriminates cases from noncases and how well the model fits the data? Explain briefly.

d. Suppose the distribution of observed and expected cases and noncases was given by the following table:

```
          Partition for the Hosmer and Lemeshow Test
                        mrsa = 1             mrsa = 0
Group  Total  Observed  Expected  Observed  Expected

 1      29       10       10.99      19       18.01
 2      31       11       10.95      20       20.05
 3      29       11       10.85      18       18.15
 4      29       11       11.73      18       17.27
 5      30       12       11.98      18       18.02
 6      31       12       11.93      19       19.07
 7      29       12       11.23      17       17.77
 8      29       13       11.42      16       17.58
 9      29       13       11.57      16       17.43
10      23        9       11.36      14       11.64
```

What does this information indicate about how well the model discriminates cases from noncases and how well the model fits the data? Explain briefly.

e. Do you think it is possible that a model might provide good discrimination between cases and noncases, yet poorly fit the data? Explain briefly, perhaps with a numerical example (e.g., using hypothetical data) or generally describing a situation, where this might happen.

Answers to Practice Exercises

1. a. $\mathbf{X}^* = (\text{CAT} = 1, \text{AGE} = 50, \text{CHL} = 200, \text{ECG} = 0, \text{SMK} = 0, \text{HPT} = 0)$

$$\hat{P}(\mathbf{X}^*) = 1/\{1 + \exp[-\text{logit } \hat{P}(\mathbf{X}^*)]\},$$

where

$$\begin{aligned}\text{logit } \hat{P}(\mathbf{X}^*) = {} & -4.0497 + (-12.6894)(1) + 0.0350(50)\\ & + (-0.00545)(200) + .3671(0) + 0.7732(0)\\ & + 1.0466(0) + 0.0692(1)(200)\\ & + (-2.3318)(1)(0)\end{aligned}$$

b. $\text{logit } \hat{P}(\mathbf{X}^*) = -2.2391$

$$\begin{aligned}\hat{P}(\mathbf{X}^*) &= 1/\{1 + \exp[-\text{logit } \hat{P}(\mathbf{X}^*)]\}\\ &= 1/\{1 + \exp[2.2391]\} = 0.096\end{aligned}$$

c. Cut-point = 0.200

Since $\hat{P}(\mathbf{X}^*) = 0.096 < 0.200$, we would predict subject $\mathbf{X}^*$ to be a noncase.

d. If the cut-point is 0 and there is at least one true case, than every case in the dataset will have $\hat{P}(\mathbf{X}^*) > 0$, i.e., all 71 true cases will exceed the cut-point and therefore be predicted to be cases. Thus, the sensitivity percent is 100(71/71) = 100.

e. It is not possible that an increase in the cut-point could result in a decrease in the specificity.

f. The denominator for computing 1 minus the specificity is the number of true noncases (538), whereas the denominator for the false positive percentage in the SAS output is the number of persons classified as positive (74). Thus, we obtain different results as follows:

Percentage specificity = (100)(1 − Sp) = (100)43/538 = 8%, whereas

Percentage false positive = (100)43/74 = 58.1%.

2. a. AUC = c = 0.789. Grade C, i.e., fair discrimination.

b. 38,198 = 71 × 538, where 71 is the number of true cases and 538 is the number of true noncases. Thus, 38,198 is the number of distinct case/noncase pairs in the dataset.

c. Percent Concordant = $100w/n_{\text{p}}$, where w is the number of case/noncase pairs for which the case has a higher predicted probability than the noncase and n_{p} is the total number of case/noncase pairs (38,198).

Percent Tied = $100z/n_{\text{p}}$, where z is the number of case/noncase pairs for which the case has the same predicted probability as the noncase.

d. $$\text{AUC} = c = \frac{w + 0.5z}{n_\text{p}}$$
$$= \frac{38{,}198(.786) + 0.5(38{,}198)(0.005)}{38{,}198} = 0.789$$

e. The AUC for the no interaction model is smaller than the AUC for the interaction model because the former model is nested within the latter model.

11 Analysis of Matched Data Using Logistic Regression

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_11, © Springer Science+Business Media, LLC 2010

Introduction

Our discussion of matching begins with a general description of the matching procedure and the basic features of matching. We then discuss how to use stratification to carry out a matched analysis. Our primary focus is on case-control studies. We then introduce the logistic model for matched data and describe the corresponding odds ratio formula. We illustrate the use of logistic regression with an application that involves matching as well as control variables not involved in matching.

We also discuss how to assess interaction involving the matching variables and whether or not matching strata should be pooled prior to analysis. Finally, we describe the logistic model for analyzing matched follow-up data.

Abbreviated Outline

The outline below gives the user a preview of this chapter. A detailed outline for review purposes follows the presentation.

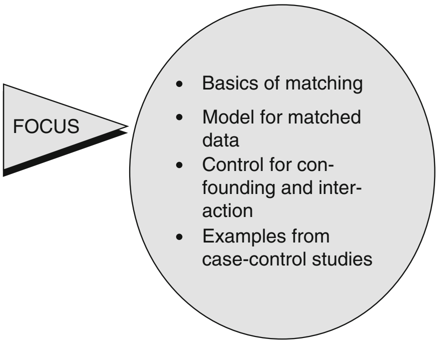
Objectives

Upon completion of this chapter, the learner should be able to:

1. State or recognize the procedure used when carrying out matching in a given study.
2. State or recognize at least one advantage and one disadvantage of matching.
3. State or recognize when to match or not to match in a given study situation.
4. State or recognize why attaining validity is not a justification for matching.
5. State or recognize two equivalent ways to analyze matched data using stratification.
6. State or recognize the McNemar approach for analyzing pair-matched data.
7. State or recognize the general form of the logistic model for analyzing matched data as an *E*, *V*, *W*-type model.
8. State or recognize an appropriate logistic model for the analysis of a specified study situation involving matched data.
9. State how dummy or indicator variables are defined and used in the logistic model for matched data.
10. Outline a recommended strategy for the analysis of matched data using logistic regression.
11. Apply the recommended strategy as part of the analysis of matched data using logistic regression.
12. Describe and/or illustrate two options for assessing interaction of the exposure variable with the matching variables in an *E*, *V*, *W*-type model.
13. Describe and/or illustrate when it would be appropriate to pool “exchangeable” matched sets.
14. State and/or illustrate the *E*, *V*, *W* model for matched follow-up data.

Presentation

I. Overview



FOCUS

- Basics of matching
- Model for matched data
- Control for confounding and interaction
- Examples from case-control studies

This presentation describes how logistic regression may be used to analyze matched data. We describe the basic features of matching and then focus on a general form of the logistic model for matched data that controls for confounding and interaction. We also provide examples of this model involving matched case-control data.

II. Basic Features of Matching

Study design procedure:

- Select referent group
- Comparable to index group on one or more "matching factors"

Matching is a procedure carried out at the design stage of a study which compares two or more groups. To match, we select a referent group for our study that is to be compared with the group of primary interest, called the index group. Matching is accomplished by constraining the referent group to be comparable to the index group on one or more risk factors, called "matching factors."

EXAMPLE

Matching factor = AGE

Referent group constrained to have *same age structure* as index group

For example, if the matching factor is age, then matching on age would constrain the referent group to have essentially the same age structure as the index group.

Case-control study:

↑

Our focus

Referent = controls

Index = cases

Follow-up study:

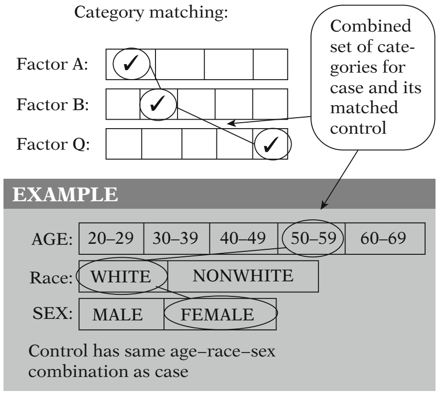
Referent = unexposed

Index = exposed

In a case-control study, the referent group consists of the controls, which is compared with an index group of cases.

In a follow-up study, the referent group consists of unexposed subjects, which is compared with the index group of exposed subjects.

Henceforth in this presentation, we focus on case-control studies, but the model and methods described apply to follow-up studies also.


Category matching:
Factor A:
Factor B:
Factor Q:
Combined set of categories for case and its matched control
EXAMPLE
AGE: 20–29 30–39 40–49 50–59 60–69
Race: WHITE NONWHITE
SEX: MALE FEMALE
Control has same age–race–sex combination as case

The most popular method for matching is called *category matching*. This involves first categorizing each of the matching factors and then finding, for each case, one or more controls from the same combined set of matching categories.

For example, if we are matching on age, race, and sex, we first categorize each of these three variables separately. For each case, we then determine his or her age–race–sex combination. For instance, the case may be 52 years old, white, and female. We then find one or more controls with the same age–race–sex combination.

Case	No. of controls	Type
1	1	1–1 or pair matching
1	R e.g., $R = 4$	R-to-1 $\longrightarrow$ 4-to-1

If our study involves matching, we must decide on the number of controls to be chosen for each case. If we decide to use only one control for each case, we call this one-to-one or pair-matching. If we choose R controls for each case, for example, R equals 4, then we call this R-to-1 matching.

R may vary from case to case

$$\text{e.g.,} \begin{cases} R = 3 \text{ for some cases} \\ R = 2 \text{ for other cases} \\ R = 1 \text{ for other cases} \end{cases}$$

Not always possible to find exactly R controls for each case

It is also possible to match so that there are different numbers of controls for different cases; that is, R may vary from case to case. For example, for some cases, there may be three controls, whereas for other cases perhaps only two or one control. This frequently happens when it is intended to do R-to-1 matching, but it is not always possible to find a full complement of R controls in the same matching category for some cases.

To match or *not to match*

As for whether to match or not in a given study, there are both advantages and disadvantages to consider.

Advantage:

Matching can be statistically *efficient*, i.e., may gain *precision* using confidence interval

The primary advantage for matching over random sampling without matching is that matching can often lead to a more statistically efficient analysis. In particular, *matching may lead to a tighter confidence interval, that is, more precision*, around the odds or risk ratio being estimated than would be achieved without matching.

Disadvantage:

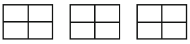
Matching is *costly*:

- To find matches
- Information loss due to discarding controls

The major disadvantage to matching is that it can be costly, both in terms of the time and labor required to find appropriate matches and in terms of information loss due to discarding of available controls not able to satisfy matching criteria. In fact, if too much information is lost from matching, it may be possible to lose statistical efficiency by matching.

Safest strategy:

Match on strong risk factors expected to be confounders

Matching	No matching
Correct estimate? YES	YES
Apropriate analysis? YES ↓ MATCHED (STRATIFIED) ANALYSIS	YES ↓ STANDARD STRATIFIED ANALYSIS
↓ SEE SECTION III	30–39 40–49 50–59  $\underbrace{\widehat{OR}_1 \quad \widehat{OR}_2 \quad \widehat{OR}_3}_{\text{combine}}$

In deciding whether to match or not on a given factor, the safest strategy is to match only on strong risk factors expected to cause confounding in the data.

Note that whether one matches or not, it is possible to obtain an unbiased estimate of the effect, namely the correct odds ratio estimate. The correct estimate can be obtained provided an appropriate analysis of the data is carried out.

If, for example, we match on age, the appropriate analysis is a *matched analysis*, which is a *special kind of stratified analysis* to be described shortly.

If, on the other hand, we do not match on age, an appropriate analysis involves dividing the data into age strata and doing a *standard stratified analysis*, which combines the results from different age strata.

Validity is not an important reason for matching (validity: getting the right answer)

Because a correct estimate can be obtained whether or not one matches at the design stage, it follows that validity is not an important reason for matching. Validity concerns getting the right answer, which can be obtained by doing the appropriate stratified analysis.

Match to gain efficiency or precision

As mentioned above, the most important statistical reason for matching is to gain efficiency or precision in estimating the odds or risk ratio of interest; that is, matching becomes worthwhile if it leads to a tighter confidence interval than would be obtained by not matching.

III. Matched Analyses Using Stratification

Strata = matched sets

The analysis of matched data can be carried out using a stratified analysis in which the strata consist of the collection of matched sets.

Special case:

Case-control study
100 matched pairs
$n = 200$
100 strata = 100 matched pairs
2 observations per stratum

As a special case, consider a pair-matched case-control study involving 100 matched pairs. The total number of observations, n, then equals 200, and the data consists of 100 strata, each of which contains the two observations in a given matched pair.

1st pair

	E	$\bar{E}$	
D			1
$\bar{D}$			1

2nd pair

	E	$\bar{E}$	
D			1
$\bar{D}$			1

. . .

100th pair

	E	$\bar{E}$	
D			1
$\bar{D}$			1

If the only variables being controlled in the analysis are those involved in the matching, then the complete data set for this matched pairs study can be represented by 100 2 × 2 tables, one for each matched pair. Each table is labeled by exposure status on one axis and disease status on the other axis. The number of observations in each table is two, one being diseased and the other (representing the control) being nondiseased.

Four possible forms:

	E	$\bar{E}$		
D	1	0	1	W pairs
$\bar{D}$	1	0	1	

	E	$\bar{E}$		
D	1	0	1	X pairs
$\bar{D}$	0	1	1	

	E	$\bar{E}$		
D	0	1	1	Y pairs
$\bar{D}$	1	0	1	

	E	$\bar{E}$		
D	0	1	1	Z pairs
$\bar{D}$	0	1	1	

Depending on the exposure status results for these data, there are four possible forms that a given stratum can take. These are shown here.

The first of these contains a matched pair for which both the case and the control are exposed.

The second of these contains a matched pair for which the case is exposed and the control is unexposed.

In the third table, the case is unexposed and the control is exposed.

And in the fourth table, both the case and the control are unexposed.

$W + X + Y + Z =$ total number of pairs

If we let W, X, Y, and Z denote the number of pairs in each of the above four types of table, respectively, then the sum W plus X plus Y plus Z equals 100, the total number of matched pairs in the study.

EXAMPLE

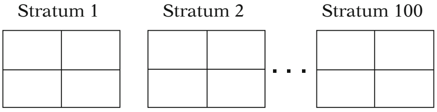
$W = 30, X = 30, Y = 10, Z = 30$

$W + X + Y + Z = 30 + 30 + 10 + 30 = 100$

For example, we may have W equals 30, X equals 30, Y equals 10, and Z equals 30, which sums to 100.

Analysis: Two equivalent ways

The analysis of a matched pair dataset can then proceed in either of two equivalent ways, which we now briefly describe.



Compute Mantel–Haenszel χ^2 and $\widehat{\text{MOR}}$

One way is to carry out a *Mantel–Haenszel chi-square test* for association based on the 100 strata and to compute a *Mantel–Haenszel odds ratio*, usually denoted as MOR, as a summary odds ratio that adjusts for the matched variables. This can be carried out using any standard computer program for stratified analysis e.g., PROC FREQUENCY, in SAS.

$\overline{D}$		D: E	D: $\overline{E}$
	E	W	Y
	$\overline{E}$	X	Z

The other method of analysis, which is equivalent to the above stratified analysis approach, is to summarize the data in a single table, as shown here. In this table, matched pairs are counted once, so that the total number of matched pairs is 100.

W:

	E	$\overline{E}$	
D	1	0	1
$\overline{D}$	1	0	1

X:

	E	$\overline{E}$	
D	1	0	1
$\overline{D}$	0	1	1

Y:

	E	$\overline{E}$	
D	0	1	1
$\overline{D}$	1	0	1

Z:

	E	$\overline{E}$	
D	0	1	1
$\overline{D}$	0	1	1

As described earlier, the quantity W represents the number of matched pairs in which both the case and the control are exposed. Similarly, X, Y, and Z are defined as previously.

$$\chi^2_{\text{MH}} = \frac{(X-Y)^2}{X+Y}, \quad \text{df} = 1$$

McNemar's test

Using the above table, the test for an overall effect of exposure, controlling for the matching variables, can be carried out using a chi-square statistic equal to the square of the difference $X - Y$ divided by the sum of X and Y. This chi-square statistic has one degree of freedom in large samples and is called *McNemar's test*.

McNemar's test = MH test for pair-matching

$\widehat{\text{MOR}} = X/Y$, 95% CI:

$$\widehat{\text{MOR}} \exp\left[\pm 196\sqrt{\tfrac{1}{X}+\tfrac{1}{Y}}\right]$$

It can be shown that McNemar's test statistic is exactly equal to the Mantel–Haenszel (MH) chi-square statistic obtained by looking at the data in 100 strata. Moreover, the MOR estimate can be calculated as X/Y, and a 95% confidence interval for the MOR can also be computed (shown on the left).

EXAMPLE

$\overline{D}$		D: E	D: $\overline{E}$
	E	$W = 30$	$Y = 10$
	$\overline{E}$	$X = 30$	$Z = 30$

$\overline{D}$		D: E	D: $\overline{E}$
	E	30	10
	$\overline{E}$	30	30

$$\chi^2_{\text{MH}} = \frac{(30-10)^2}{30+10} = \frac{400}{40} = 10.0$$

As an example of McNemar's test, suppose W equals 30, X equals 30, Y equals 10, and Z equals 30, as shown in the table here.

Then based on these data, the McNemar test statistic is computed as the square of 30 minus 10 divided by 30 plus 10, which equals 400 over 40, which equals 10.

EXAMPLE (continued)

χ^2 ~chi square 1 df
under H_0: OR = 1

$P << 0.01$, significant

$$\widehat{\text{MOR}} = \frac{X}{Y} = 3, \quad 95\%\ \text{CI}: \ (2.31, 6.14)$$

This statistic has approximately a chi-square distribution with one degree of freedom under the null hypothesis that the odds ratio relating exposure to disease equals 1.

From chi-square tables, we find this statistic to be highly significant with a *P*-value well below 0.01.

The estimated odds ratio, which adjusts for the matching variables, can be computed from the above table using the MOR formula *X* over *Y* which in this case turns out to be 3. The computed 95% confidence interval is also shown at the left.

Analysis for *R*-to-1 and mixed matching use stratified analysis

We have thus described how to do a matched pair analysis using stratified analysis or an equivalent McNemar's procedure. If the matching is *R*-to-1 or even involves mixed matching ratios, the analysis can also be done using a stratified analysis.

EXAMPLE

$R = 4$: Illustrating one stratum

	E	$\overline{E}$	
D	1	0	1
$\overline{D}$	1	3	4
			5

For example, if *R* equals 4, then each stratum contains five subjects, consisting of the one case and its four controls. These numbers can be seen on the margins of the table shown here. The numbers inside the table describe the numbers exposed and unexposed within each disease category. Here, we illustrate that the case is exposed and that three of the four controls are unexposed. The breakdown within the table may differ with different matched sets.

R-to-1 or mixed matching

use χ^2_{MH} and $\widehat{\text{MOR}}$
for stratified data

Nevertheless, the analysis for *R*-to-1 or mixed matched data can proceed as with pair-matching by computing a Mantel–Haenszel chi-square statistic and a Mantel–Haenszel odds ratio estimate based on the stratified data.

IV. The Logistic Model for Matched Data

1. Stratified analysis
2. McNemar analysis

✓3. Logistic modeling

A third approach to carrying out the analysis of matched data involves logistic regression modeling.

Advantage of modeling
can control for variables *other* than matched variables

The main advantage of using logistic regression with matched data occurs when there are variables other than the matched variables that the investigator wishes to control.

EXAMPLE

Match on AGE, RACE, SEX
also, control for SBP and BODYSIZE

For example, one may match on AGE, RACE, and SEX, but may also wish to control for systolic blood pressure and body size, which may have also been measured but were not part of the matching.

Logistic model for matched data includes control of variables not matched

In the remainder of the presentation, we describe how to formulate and apply a logistic model to analyze matched data, which allows for the control of variables not involved in the matching.

Stratified analysis inefficient:
Data is discarded

In this situation, using a stratified analysis approach instead of logistic regression will usually be inefficient in that much of one's data will need to be discarded, which is not required using a modeling approach.

Matched data:
Use conditional ML estimation (number of parameters large relative to n)

The model that we describe below for matched data requires the use of conditional ML estimation for estimating parameters. This is because, as we shall see, when there are matched data, the number of parameters in the model is large relative to the number of observations.

Pair-matching:

$$\underset{\text{overestimate}}{\underset{\uparrow}{\widehat{\text{OR}}_{\text{U}}}} = (\widehat{\text{OR}}_{\text{C}})^2$$

If unconditional ML estimation is used instead of conditional, an overestimate will be obtained. In particular, for pair-matching, the estimated odds ratio using the unconditional approach will be the square of the estimated odds ratio obtained from the conditional approach, the latter being the correct result.

Principle
Matched analysis ⇒ stratified analysis

- Strata are matched sets, e.g., pairs
- Strata defined using dummy (indicator) variables

An important principle about modeling matched data is that such modeling requires the matched data to be considered in strata. As described earlier, the strata are the matched sets, for example, the pairs in a matched pair design. In particular, the strata are defined using *dummy* or indicator variables, which we will illustrate shortly.

$E = (0, 1)$ exposure

$C_1, C_2, \ldots, C_p$ control variables

In defining a model for a matched analysis, we consider the special case of a single (0, 1) exposure variable of primary interest, together with a collection of control variables C_1, C_2, and so on up through C_p, to be adjusted in the analysis for possible confounding and interaction effects.

- Some Cs matched by design
- Remaining Cs not matched

We assume that some of these C variables have been matched in the study design, either using pair-matching or R-to-1 matching. The remaining C variables have not been matched, but it is of interest to control for them, nevertheless.

$D = (0, 1)$ disease

$X_1 = E = (0, 1)$ exposure

Given the above context, we now define the following set of variables to be incorporated into a logistic model for matched data. We have a (0, 1) disease variable D and a (0, 1) exposure variable X_1 equal to E.

Some Xs: V_{1i} dummy variables (matched strata)

We also have a collection of Xs which are dummy variables to indicate the different matched strata; these variables are denoted as V_1 variables.

Some Xs: V_{2j} variables (potential confounders)

Further, we have a collection of Xs which are defined from the Cs not involved in the matching and represent potential confounders in addition to the matched variables. These potential confounders are denoted as V_2 variables.

Some Xs: product terms EW_j
(Note: Ws usually V_2s)

And finally, we have a collection of Xs which are product terms of the form E times W, where the Ws denote potential interaction variables. Note that the Ws will usually be defined in terms of the V_2 variables.

The model:

$$\begin{aligned} \text{logit P}(\mathbf{X}) = {} & \alpha + \beta E \\ & + \sum \underbrace{\gamma_{1i} V_{1i}}_{\text{matching}} + \sum \underbrace{\gamma_{2j} V_{2j}}_{\text{confounders}} \\ & + E \sum \underbrace{\delta_k W_k}_{\text{interaction}} \end{aligned}$$

The logistic model for matched analysis is then given in logit form as shown here. In this model, the γ_{1i}s are coefficients of the dummy variables for the matching strata, the γ_{2i}s are the coefficients of the potential confounders not involved in the matching, and the δ_js are the coefficients of the interaction variables.

EXAMPLE

Pair-matching by AGE, RACE, SEX
100 matched pairs
99 dummy variables

$$V_{1i} = \begin{cases} 1 & \text{if } i\text{th matched pair} \\ 0 & \text{otherwise} \end{cases}$$

$$i = 1, 2, \ldots, 99$$

$$\text{V}_{11} = \begin{cases} 1 & \text{if first matched pair} \\ 0 & \text{otherwise} \end{cases}$$

$$V_{12} = \begin{cases} 1 & \text{if second matched pair} \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

$$V_{1,99} = \begin{cases} 1 & \text{if 99th matched pair} \\ 0 & \text{otherwise} \end{cases}$$

As an example of dummy variables defined for matched strata, consider a study involving pair-matching by AGE, RACE, and SEX, containing 100 matched pairs. Then, the above model requires defining 99 dummy variables to incorporate the 100 matched pairs.

We can define these dummy variables as V_{1i} equals 1 if an individual falls into the ith matched pair and 0 otherwise. Thus, it follows that V_{11} equals 1 if an individual is in the first matched pair and 0 otherwise, V_{12} equals 1 if an individual is in the second matched pair and 0 otherwise, and so on up to $V_{1,\,99}$, which equals 1 if an individual is in the 99th matched pair and 0 otherwise.

EXAMPLE (continued)

1st matched set

$V_{11} = 1, V_{12} = V_{13} = \cdots = V_{1,99} = 0$

99th matched set

$V_{1,99} = 1, V_{11} = V_{12} = \cdots = V_{1,98} = 0$

100th matched set

$V_{11} = V_{12} = \cdots = V_{1,99} = 0$

Alternatively, using the above dummy variable definition, a person in the first matched set will have V_{11} equal to 1 and the remaining dummy variables equal to 0; a person in the 99th matched set will have $V_{1,\ 99}$ equal to 1 and the other dummy variables equal to 0; and a person in the 100th matched set will have all 99 dummy variables equal to 0.

Matched pairs model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \sum \gamma_{1i} V_{1i} + \sum \gamma_{2j} V_{2j} + E \sum \delta_k W_k$$

$$\text{ROR} = \exp\left(\beta + \sum \delta_k W_k\right)$$

For the matched analysis model we have just described, the odds ratio formula for the effect of exposure status adjusted for covariates is given by the expression ROR equals e to the quantity β plus the sum of the δ_j times the W_j.

Note: Two types of V variables are controlled

This is exactly the same odds ratio formula given in our review for the E, V, W model. This makes sense because the matched analysis model is essentially an E, V, W model containing two different types of V variables.

V. An Application

EXAMPLE

Case-control study
2-to-1 matching

$D = \text{MI}_{0,1}$

$E = \text{SMK}_{0,1}$

$$\underbrace{C_1 = \text{AGE}, C_2 = \text{RACE}, C_3 = \text{SEX}, C_4 = \text{HOSPITAL}}_{\text{matched}}$$

$$\underbrace{C_5 = \text{SBP}\ C_6 = \text{ECG}}_{\text{not matched}}$$

As an application of a matched pairs analysis, consider a case-control study involving 2-to-1 matching which involves the following variables:

The *disease variable* is myocardial infarction status, as denoted by MI.

The *exposure variable* is smoking status, as defined by a (0, 1) variable denoted as SMK.

There are six C variables to be controlled. The first four of these variables, namely age, race, sex, and hospital status, are involved in the matching.

The last two variables, systolic blood pressure, denoted by SBP, and electrocardiogram status, denoted by ECG, are not involved in the matching.

EXAMPLE (continued)

$n = 117$ (39 matched sets)

The study involves 117 persons in 39 matched sets, or strata, each strata containing 3 persons, 1 of whom is a case and the other 2 are matched controls.

The model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta\text{SMK} + \sum_{i=1}^{38} \gamma_{1i}V_{1i}$$

$$= \underbrace{\gamma_{21}\,\text{SBP} + \gamma_{22}\,\text{ECG}}_{\text{confounders}}$$

$$+ \text{SMK}\,\underbrace{(\delta_1\text{SBP} + \delta_2\text{ECG})}_{\text{modifiers}}$$

The logistic model for the above situation can be defined as follows: logit P(**X**) equals α plus β times SMK plus the sum of 38 terms of the form γ_{1i} times V_{1i}, where V_{1i}s are dummy variables for the 39 matched sets, plus γ_{21} times SBP plus γ_{22} times ECG plus SMK times the sum of δ_1 times SBP plus δ_2 times ECG.

Here, we are considering two potential confounders involving the two variables (SBP and ECG) not involved in the matching and also two interaction variables involving these same two variables.

$$\text{ROR} = \exp(\beta + \delta_1\text{SBP} + \delta_2\text{ECG})$$

The odds ratio for the above logistic model is given by the formula e to the quantity β plus the sum of δ_1 times SBP and δ_2 times ECG.

β = coefficient of E
δ_1 = coefficient of E × SBP
δ_2 = coefficient of E × ECG

Note that this odds ratio expression involves the coefficients β, δ_1, and δ_2, which are coefficients of variables involving the exposure variable. In particular, δ_1 and δ_2 are coefficients of the interaction terms E × SBP and E × ECG.

$$\text{Starting model} \xrightarrow{\text{analysis strategy}} \text{Final model}$$

The model we have just described is the starting model for the analysis of the dataset on 117 subjects. We now address how to carry out an analysis strategy for obtaining a final model that includes only the most relevant of the covariates being considered initially.

Estimation method:
✓ Conditional ML estimation
(also, we illustrate unconditional ML estimation)

The first important issue in the analysis concerns the choice of estimation method for obtaining ML estimates. Because matching is being used, the appropriate method is conditional ML estimation. Nevertheless, we also show the results of unconditional ML estimation to illustrate the type of bias that can result from using the wrong estimation method.

Interaction:
SMK × SBP and SMK × ECG?

The next issue to be considered is the assessment of interaction. Based on our starting model, we, therefore, determine whether or not either or both of the product terms SMK × SBP and SMK × ECG are retained in the model.

EXAMPLE (continued)

Chunk test:

$H_0: \delta_1 = \delta_2 = 0$,

where

$\delta_1 =$ coefficient of SMK × SBP

$\delta_2 =$ coefficient of SMK × ECG

One way to test for this interaction is to carry out a chunk test for the significance of both product terms considered collectively. This involves testing the null hypothesis that the coefficients of these variables, namely δ_1 and δ_2, are both equal to 0.

$$\text{LR} = \left(-2\ln \hat{L}_\text{R}\right) - \left(-2\ \ln\ \hat{L}_\text{F}\right)$$

R = reduced model (no interaction) F = full model (interaction)

Log likelihood statistics

$-2\ \ln\ \hat{L}$

The test statistic for this chunk test is given by the likelihood ratio (LR) statistic computed as the difference between log likelihood statistics for the full model containing both interaction terms and a reduced model which excludes both interaction terms. The log likelihood statistics are of the form -2 ln $\hat{L}$, where $\hat{L}$ is the maximized likelihood for a given model.

$\text{LR} \sim \chi^2_2$

Number of parameters tested = 2

This likelihood ratio statistic has a chi-square distribution with two degrees of freedom. The degrees of freedom are the number of parameters tested, namely 2.

$-2\ \ln\ \hat{L}_\text{F} = 60.23$

$-2\ \ln\ \hat{L}_\text{R} = 60.63$

When carrying out this test, the log likelihood statistics for the full and reduced models turn out to be 60.23 and 60.63, respectively.

$\text{LR} = 60.63 - 60.23 = 0.40$

$P > 0.10$ (no significant interaction)

Therefore, drop SMK × SBP and SMK × ECG from model

The difference between these statistics is 0.40. Using chi-square tables with two degrees of freedom, the P-value is considerably larger than 0.10, so we can conclude that there are no significant interaction effects. We can, therefore, drop the two interaction terms from the model.

Backward elimination: same conclusion

Note that an alternative approach to testing for interaction is to use backward elimination on the interaction terms in the initial model. Using this latter approach, it turns out that both interaction terms are eliminated. This strengthens the conclusion of no interaction.

$$\text{logit P}(\mathbf{X}) = \alpha + \beta\text{SMK} + \sum \gamma_{1i}V_{1i} + \gamma_{21}\text{SBP} + \gamma_{22}\text{ECG}$$

At this point, our model can be simplified to the one shown here, which contains only main effect terms. This model contains the exposure variable SMK, 38 V variables that incorporate the 39 matching strata, and 2 V variables that consider the potential confounding effects of SBP and ECG, respectively.

EXAMPLE (continued)

$\widehat{\text{ROR}} = e^{\hat{\beta}}$

Under this reduced model, the estimated odds ratio adjusted for the effects of the *V* variables is given by the familiar expression e to the $\hat{\beta}$, where $\hat{\beta}$ is the coefficient of the exposure variable SMK.

*V*s in model	OR = e^{β}	95% CI
SBP and ECG	*C* $\boxed{2.07}$ *U* 3.38	(0.69, 6.23)
SBP only	*C* 2.08 *U* 3.39	(0.72, 6.00)
ECG only	*C* 2.05 *U* 3.05	(0.77, 5.49)
Neither	*C* 2.32 *U* 3.71	(0.93, 5.79)

C = conditional estimate
U = unconditional estimate

The results from fitting this model and reduced versions of this model which delete either or both of the potential confounders SBP and ECG are shown here. These results give both conditional (*C*) and unconditional (*U*) odds ratio estimates and 95% confidence intervals (CI) for the conditional estimates only. (See Computer Appendix.)

From inspection of this table of results, we see that the unconditional estimation procedure leads to overestimation of the odds ratio and, therefore, should not be used.

Minimal confounding:
Gold standard $\widehat{\text{OR}} = 2.07$, essentially same as other $\widehat{\text{OR}}$

The results also indicate a minimal amount of confounding due to SBP and ECG. This can be seen by noting that the gold standard estimated odds ratio of 2.07, which controls for both SBP and ECG, is essentially the same as the other conditionally estimated odds ratios that control for either SBP or ECG or neither.

But 2.07 moderately different from 2.32, so we control for *at least* one of SBP and ECG

Nevertheless, because the estimated odds ratio of 2.32, which ignores both SBP and ECG in the model, is moderately different from 2.07, we recommend that at least one or possibly both of these variables be controlled.

Narrowest CI: Control for ECG only

If at least one of SBP and ECG is controlled, and confidence intervals are compared, the narrowest confidence interval is obtained when only ECG is controlled.

Most precise estimate:
Control for ECG only

Thus, the most precise estimate of the effect is obtained when ECG is controlled, along, of course, with the matching variables.

All CI are wide and include 1

Overall conclusion:
Adjusted $\widehat{\text{OR}} \approx 2$, but is nonsignificant

Nevertheless, because all confidence intervals are quite wide and include the null value of 1, it does not really matter which variables are controlled. The overall conclusion from this analysis is that the adjusted estimate of the odds ratio for the effect of smoking on the development of MI is about 2, but it is quite nonsignificant.

VI. Assessing Interaction Involving Matching Variables

EXAMPLE

$D = \text{MI}$

$E = \text{SMK}$

AGE, RACE, SEX, HOSPITAL: matched

SBP, ECG: not matched

The previous section considered a study of the relationship between smoking (SMK) and myrocardial infarction (MI) in which cases and controls were matched on four variables: AGE, RACE, SEX, and Hospital. Two additional control variables, SBP and ECG, were not involved in the matching.

Interaction terms:

SMK × SBP, SMK × ECG

tested using LR test

In the above example, interaction was evaluated by including SBP and ECG in the logistic regression model as product terms with the exposure variable SMK. A test for interaction was then carried out using a likelihood ratio test to determine whether these two product terms could be dropped from the model.

Interaction between

SMK and matching variables?

Two options.

Suppose the investigator is also interested in considering possible interaction between exposure (SMK) and one or more of the matching variables. The proper approach to take in such a situation is not as clear-cut as for the previous interaction assessment. We now discuss two options for addressing this problem.

Option 1:
Add product terms of the form

$$E \times V_{1i}$$

The first option involves adding product terms of the form $E \times V_{1i}$ to the model for each dummy variable V_{1i} indicating a matching stratum.

$$\text{logit}\, P(\mathbf{X}) = \alpha + \beta E + \sum_i \gamma_{1i} V_{1i} + \sum_i \gamma_{2j} V_{2j} + E \sum_i \delta_{1i} V_{1i} + E \sum_k \delta_k W_k,$$

where

V_{1i} = dummy variables for matching strata

V_{2j} = other covariates (not matched)

W_k = effect modifiers defined from other covariates

The general form of the logistic model that accommodates interaction defined using this option is shown on the left. The expression to the right of the equals sign includes terms for the intercept, the main exposure (i.e., SMK), the matching strata, other control variables not matched on, product terms between the exposure and the matching strata, and product terms between the exposure and other control variables not matched on.

EXAMPLE (continued)

Option 1:

Test H_0: All $\delta_{1i} = 0$.
(Chunk test)

Not significant ⇒ No interaction involving matching variables

Significant ⇒ Interaction involving matching variables

⇒ Carry out backward elimination of $E \times V_{1i}$ terms

Using the above (option 1) interaction model, we can assess interaction of exposure with the matching variables by testing the null hypothesis that all the coefficients of the $E \times V_{1i}$ terms (i.e., all the δ_{1i}) are equal to zero.

If this “chunk” test is not significant, we could conclude that there is no interaction involving the matching variables. If the test is significant, we might then carry out backward elimination to determine which of the $E \times V_{1i}$ terms need to stay in the model. (We could also carry out backward elimination even if the “chunk” test is nonsignificant.)

Criticisms of option 1:

- Difficult to determine which of several matching variables are effect modifiers. (The V_{1i} represent matching strata, not matching variables.)

A criticism of this (option 1) approach is that if significant interaction is found, then it will be difficult to determine which of possibly several matching variables are effect modifiers. This is because the dummy variables (V_{1i}) in the model represent matching strata rather than specific effect modifier variables.

- Not enough data to assess interaction (number of parameters may exceed n).

Another problem with option 1 is that there may not be enough data in each stratum (e.g., when pair-matching) to assess interaction. In fact, if there are more parameters in the model than there are observations in the study, the model will not execute.

Option 2:
Add product terms of the form

$$E \times W_{1\mathrm{m}},$$

where $W_{1\mathrm{m}}$ is a *matching variable*

$$\operatorname{logit} \mathrm{P}(\mathrm{X}) = \alpha+\beta E=\sum_i \gamma_{1i}V_{1i}+\sum_j \gamma_{2j}V_{2j} +E\sum_{\mathrm{m}} \delta_{1\mathrm{m}}W_{1\mathrm{m}} +E\sum_k \delta_k W_k,$$

where
$W_{1\mathrm{m}}$ = matching variables in original form
W_{2k} = effect modifiers defined from other covariates (not matched)

A second option for assessing interaction involving matching variables is to consider product terms of the form $E \times W_{1\mathrm{m}}$, where $W_{1\mathrm{m}}$ is an actual matching variable.

The corresponding logistic model is shown at the left. This model contains the exposure variable E, dummy variables V_{1i} for the matching strata, nonmatched covariates V_{2j}, product terms $E \times W_{1\mathrm{m}}$ involving the matching variables, and $E \times W_k$ terms, where the W_k are effect modifiers defined from the unmatched covariates.

EXAMPLE (continued)

Option 2:

Test H_0: All $\delta_{1m} = 0$.
(Chunk test)

Not significant ⇒ No interaction involving matching variables

Significant ⇒ Interaction involving matching variables
⇒ Carry out Backwards Elimination of $E \times W_{1m}$ terms

Using the above (option 2) interaction model, we can assess interaction of exposure with the matching variables by testing the null hypothesis that all of the coefficients of the $E \times W_{1m}$ terms (i.e., all of the δ_{1m}) equal zero.

As with option 1, if the "chunk" test for interaction involving the matching variables is not significant, we could conclude that there is no interaction involving the matching variables. If, however, the chunk test is significant, we might then carry out backward elimination to determine which of the $E \times W_{1m}$ terms should remain in the model. We could also carry out backward elimination even if the chunk test is not significant.

Criticism of option 2:

The model is technically not HWF.

$E \times W_{1m}$ in model but not W_{1m}

A problem with the second option is that the model for this option is not hierarchically well-formulated (HWF), since components (W_{1m}) of product terms ($E \times W_{1m}$) involving the matching variables are not in the model as main effects. (See Chap. 6 for a discussion of the HWF criterion.)

	Option 1	Option 2
Interpretable?	No	Yes
HWF?	Yes	No (but almost yes)

Although both options for assessing interaction involving matching variables have problems, the second option, though not HWF, allows for a more interpretable decision about which of the matching variables might be effect modifiers. Also, even though the model for option 2 is technically not HWF, the matching variables are at least in some sense in the model as both effect modifiers and confounders.

Alternatives to options 1 and 2:

- Do not match on any variable that you consider a possible effect modifier.
- Do not assess interaction for any variable that you have matched on.

One way to avoid having to choose between these two options is to decide not to match on any variable that you wish to assess as an effect modifier. Another alternative is to avoid assessing interaction involving any of the matching variables, which is often what is done in practice.

VII. Pooling Matching Strata

To pool or not to pool matched sets?

Another issue to be considered in the analysis of matched data is whether to combine, or *pool*, matched sets that have the same values for all variables being matched on.

Case-control study:

- Pair-match on SMK (ever vs. never)
- 100 cases (i.e., $n = 200$)
- Smokers – 60 matched pairs
- Nonsmokers – 40 matched pairs

Suppose smoking status (SMK), defined as ever vs. never smoked, is the only matching variable in a pair-matched case-control study involving 100 cases. Suppose further that when the matching is carried out, 60 of the matched pairs are all smokers and the 40 remaining matched pairs are all nonsmokers.

Matched pair A	Matched pair B
Case A – Smoker	Case B – Smoker
Control A – Smoker ↔	Control B – Smoker

(interchangeable)

Now, let us consider any two of the matched pairs involving smokers, say pair A and pair B. Since the only variable being matched on is smoking, the control in pair A had been eligible to be chosen as the control for the case in pair B prior to the matching process. Similarly, the control smoker in pair B had been eligible to be the control smoker for the case in pair A.

Controls for matched pairs A and B are interchangeable

⇓

Matched pairs A and B are *exchangeable* (definition)

Even though this did not actually happen after matching took place, the potential interchangeability of these two controls suggests that pairs A and B should not be treated as separate strata in a matched analysis. Matched sets such as pairs A and B are called *exchangeable* matched sets.

Smokers: 60 matched pairs are exchangeable

Nonsmokers: 40 matched pairs are exchangeable

For the entire study involving 100 matched pairs, the 60 matched pairs all of whom are smokers are exchangeable and the remaining 40 matched pairs of nonsmokers are separately exchangeable.

Ignoring exchangeability

⇓

Use stratified analysis with 100 strata, e.g., McNemar's test

If we ignored exchangeability, the typical analysis of these data would be a stratified analysis that treats all 100 matched pairs as 100 separate strata. The analysis could then be carried out using the discordant pairs information in McNemar's table, as we described in Sect. III.

Ignore exchangeability? *No!!!*

Treating such strata separately is artificial,

i.e., exchangeable strata are not unique

But should we actually ignore the exchangeability of matched sets? We say no, primarily because to treat exchangeable strata separately artificially assumes that such strata are unique from each other when, in fact, they are not. [In statistical terms, we argue that adding parameters (e.g., strata) unnecessarily to a model results in a loss of precision.]

Analysis? Pool exchangeable matched sets

How should the analysis be carried out? The answer here is to pool exchangeable matched sets.

EXAMPLE (match on SMK)

Use two pooled strata:

Stratum 1: Smokers ($n = 60 \times 2$)

Stratum 2: Nonsmokers ($n = 40 \times 2$)

In our example, pooling would mean that rather than analyzing 100 distinct strata with 2 persons per strata, the analysis would consider only 2 pooled strata, one pooling 60 matched sets into a smoker's stratum and the other pooling the other 40 matched sets into a nonsmoker's stratum.

Matching on several variables

⇓

May be only a few exchangeable matched sets

⇓

Pooling has negligible effect on odds ratio estimates

More generally, if several variables are involved in the matching, the study data may only contain a relatively low number of exchangeable matched sets. In such a situation, the use of a pooled analysis, even if appropriate, is likely to have a negligible effect on the estimated odds ratios and their associated standard errors, when compared with an unpooled matched analysis.

However, pooling may greatly reduce the number of strata to be analyzed (e.g., from 100 to 2 strata)

It is, nevertheless, quite possible that the pooling of exchangeable matched sets may greatly reduce the number of strata to be analyzed. For example, in the example described earlier, in which smoking was the only variable being matched, the number of strata was reduced from 100 to only 2.

If no. of strata greatly reduced by pooling

⇓

Unconditional ML may be used if "appropriate"

When pooling reduces the number of strata considerably, as in the above example, it may then be appropriate to use an unconditional maximum likelihood procedure to fit a logistic model to the pooled data.

EXAMPLE (continued)

Unconditional ML estimation ‘‘appropriate’’ provided

$\text{OR}_{\text{unconditional}}$ *unbiased*

and

$\text{CI}_{\text{unconditional}}$ *narrower than* $\text{CI}_{\text{conditional}}$

By “appropriate,” we mean that the odds ratio from the unconditional ML approach should be unbiased, and may also yield a narrower confidence interval around the odds ratio. Conditional ML estimation will always give an unbiased estimate of the odds ratio, however.

Summary on pooling:

Recommend:

- Identify and pool exchangeable matched sets
- Carry out stratified analysis or logistic regression using pooled strata
- Consider using unconditional ML estimation (but conditional ML estimation always unbiased)

To summarize our discussion of pooling, we recommend that whenever matching is used, the investigator should identify and pool exchangeable matched sets. The analysis can then be carried out using the reduced number of strata resulting from pooling using either a stratified analysis or logistic regression. If the resulting number of strata is small enough, then unconditional ML estimation may be appropriate. Nevertheless, conditional ML estimation will always ensure that estimated odds ratios are unbiased.

VIII. Analysis of Matched Follow-up Data

Follow-up data:
Unexposed = referent
Exposed = index

Thus far we have considered only matched case-control data. We now focus on the analysis of matched cohort data.

Unexposed and exposed groups have same distribution of matching variables.

In follow-up studies, matching involves the selection of unexposed subjects (i.e., the referent group) to have the same or similar distribution as exposed subjects (i.e., the index group) on the matching variables.

	Exposed	Unexposed
White male	30%	30%
White female	20%	20%
Nonwhite male	15%	15%
Nonwhite female	35%	35%

If, for example, we match on race and sex in a follow-up study, then the unexposed and exposed groups should have the same/similar race by sex (combined) distribution.

Individual matching
or
Frequency matching (more convenient, larger sample size)

As with case-control studies, matching in follow-up studies may involve either individual matching (e.g., R-to-1 matching) or frequency matching. The latter is more typically used because it is convenient to carry out in practice and allows for a larger total sample size once a cohort population has been identified.

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \sum_i \gamma_{1i} V_{1i} + \sum_j \gamma_{2j} V_{2j} + E \sum_k \delta_k W_k,$$

where

V_{1i} = dummy variables for matching strata

V_{2j} = other covariates (not matched)

W_k = effect modifiers defined from other covariates

The logistic model for matched follow-up studies is shown at the left. This model is essentially the same model as we defined for case-control studies, except that the matching strata are now defined by exposed/unexposed matched sets instead of by case/control matched sets. The model shown here allows for interaction between the exposure of interest and the control variables that are not involved in the matching.

Frequency matching (small no. of strata)

$\Downarrow$

Unconditional ML estimation may be used if "appropriate" (Conditional ML always unbiased})

If frequency matching is used, then the number of matching strata will typically be small relative to the total sample size, so it is appropriate to consider using unconditional ML estimation for fitting the model. Nevertheless, as when pooling exchangeable matched sets results from individual matching, conditional ML estimation will always provide unbiased estimates (but may yield less precise estimates than obtained from unconditional ML estimation).

Four types of stratum:

Type 1

	E	$\bar{E}$
D	1	1
$\bar{D}$	0	0
	1	1

P pairs concordant

Type 2

	E	$\bar{E}$
D	1	0
$\bar{D}$	0	1
	1	1

Q pairs discordant

Type 3

	E	$\bar{E}$
D	0	1
$\bar{D}$	1	0
	1	1

R pairs discordant

Type 4

	E	$\bar{E}$
D	0	0
$\bar{D}$	1	1
	1	1

S pairs concordant

In matched-pair follow-up studies, each of the matched sets (i.e., strata) can take one of four types, shown at the left. This is analogous to the four types of stratum for a matched case-control study, except here each stratum contains one exposed subject and one unexposed subject rather than one case and control.

The first of the four types of stratum describes a "concordant" pair for which both the exposed and unexposed have the disease. We assume there are P pairs of this type.

The second type describes a "discordant pair" in which the exposed subject is diseased and an unexposed subject is not diseased. We assume Q pairs of this type.

The third type describes a "discordant pair" in which the exposed subject is nondiseased and the unexposed subject is diseased. We assume R pairs of this type.

The fourth type describes a “concordant pair” in which both the exposed and the unexposed do not have the disease. Assume S pairs of this type.

Stratified analysis:

Each matched pair is a stratum

or

Pool exchangeable matched sets

The analysis of data from a matched pair follow-up study can then proceed using a stratified analysis in which each matched pair is a separate stratum or the number of strata is reduced by pooling exchangeable matched sets.

$$\begin{array}{cc|cc|} & & \multicolumn{2}{c}{\bar{E}} \\ & & D & \bar{D} \\ \cline{3-4} E & D & P & Q \\ & \bar{D} & R & S \\ \cline{3-4} \end{array}$$

Without pooling → McNemar's table

$$\widehat{\text{MRR}} = \frac{P+Q}{P+R} \qquad \widehat{\text{MOR}} = \frac{Q}{R}$$

$$\chi^2_{\text{MH}} = \frac{(Q-R)^2}{Q+R}$$

If pooling is not used, then, as with case-control matching, the data can be rearranged into a McNemar-type table as shown at the left. From this table, a Mantel–Haenszel risk ratio can be computed as $(P + Q)/(P + R)$. Also, a Mantel–Haenszel odds ratio is computed as Q/R.

Furthermore, a Mantel–Haenszel test of association between exposure and disease that controls for the matching is given by the chi-square statistic $(Q - R)^2/(Q + R)$, which has one degree of freedom under the null hypothesis of no E–D association.

$\widehat{\text{MOR}}$ and χ^2_{MH} use discordant pairs information

$\widehat{\text{MRR}}$ uses discordant and concordant pairs information

In the formulas described above, both the Mantel–Haenszel test and odds ratio estimate involve only the discordant pair information in the McNemar table. However, the Mantel–Haenszel risk ratio formula involves the concordant diseased pairs in addition to the discordant pairs.

EXAMPLE

Pair-matched follow-up study 4,830 matched pairs

E = VS $(0 = \text{no},\ 1 = \text{yes})$

D = MI $(0 = \text{no},\ 1 = \text{yes})$

Matching variables: AGE and YEAR

As an example, consider a pair-matched follow-up study with 4,830 matched pairs designed to assess whether vasectomy is a risk factor for myocardial infarction. The exposure variable of interest is vasectomy status (VS: 0 = no, 1 = yes), the disease is myocardial infarction (MI: 0 = no, 1 = yes), and the matching variables are AGE and YEAR (i.e., calendar year of follow-up).

EXAMPLE (continued)

McNemar's table:

		VS = 0	
		MI = 1	MI = 0
VS = 1	MI = 1	P = 0	Q = 20
	MI = 0	R = 16	S = 4790

$$\widehat{\text{MRR}} = \frac{P+Q}{P+R} = \frac{0+20}{0+16} = 1.25$$

Note: $P = 0 \Rightarrow \widehat{\text{MRR}} = \widehat{\text{MOR}}$.

$$\chi^2_{\text{MH}} = \frac{(Q-R)^2}{Q+R} = \frac{(20-16)^2}{20+16} = 0.44$$

Cannot reject H_0: mRR = 1

If no other covariates are considered other than the matching variables (and the exposure), the data can be summarized in the McNemar table shown at the left.

From this table, the estimated MRR, which adjusts for AGE and YEAR equals 20/16 or 1.25. Notice that since $P = 0$ in this table, the $\widehat{\text{MRR}}$ equals the $\widehat{\text{MOR}} = Q/R$.

The McNemar test statistic for these data is computed to be $\chi^2_{\text{MH}} = 0.44$ (df = 1), which is highly nonsignificant. Thus, from this analysis we cannot reject the null hypothesis that the risk ratio relating vasectomy to myocardial infarction is equal to its null value (i.e., 1).

Criticism:

- Information on 4,790 discordant pairs not used

The analysis just described could be criticized in a number of ways. First, since the analysis only used the 36 discordant pairs information, all of the information on the 4,790 concordant pairs was not needed, other than to distinguish such pairs from concordant pairs.

- Pooling exchangeable matched sets more appropriate analysis

Second, since matching involved only two variables, AGE and YEAR, a more appropriate analysis should have involved a stratified analysis based on pooling exchangeable matched sets.

- Frequency matching more appropriate than individual matching

Third, a more appropriate design would likely have used frequency matching on AGE and YEAR rather than individual matching.

How to modify the analysis to control for nonmatched variables

OBS and SMK?

Assuming that a more appropriate analysis would have arrived at essentially the same conclusion (i.e., a negative finding), we now consider how the McNemar analysis described above would have to be modified to take into account two additional variables that were not involved in the matching, namely obesity status (OBS) and smoking status (SMK).

Matched + nonmatched variables

$\Downarrow$

Use logistic regression

When variables not involved in the matching, such as OBS and SMK, are to be controlled in addition to the matching variable, we need to use logistic regression analysis rather than a stratified analysis based on a McNemar data layout.

No interaction model:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta \text{VS} + \sum_i \gamma_{1i} \text{V}_{1i} + \gamma_{21}\text{OBS} + \gamma_{22}\text{SMK}$$

A no-interaction logistic model that would accomplish such an analysis is shown at the left. This model takes into account the exposure variable of interest (i.e., VS) as well as the two variables not matched on (i.e., OBS and SMK), and also includes terms to distinguish the different matched pairs (i.e., the V_{1i} variables).

4830 total pairs ↔ 36 discordant pairs

same results

It turns out (from statistical theory) that the results from fitting the above model would be identical regardless of whether all 4,380 matched pairs or just the 36 discordant matched pairs are input as the data.

Need only analyze discordant pairs

In other words, for pair-matched follow-up studies, even if variables not involved in the matching are being controlled, a logistic regression analysis requires only the information on discordant pairs to obtain correct estimates and tests.

Pair-matched case-control studies:

Use only discordant pairs

provided

no other control variables other than matching variables

The above property of pair-matched follow-up studies does NOT hold for pair-matched case-control studies. For the latter, discordant pairs should only be used if there are no other control variables other than the matching variables to be considered in the analysis. In other words, for pair-matched case-control data, if there are unmatched variables being controlled, the complete dataset must be used to obtain correct results.

IX. SUMMARY

This presentation:

- Basic features of matching
- Logistic model for matched data
- Illustration using 2-to-1 matching
- Interaction involving matching variables
- Pooling exchangeable matched sets
- Matched follow-up data

This presentation is now complete. In summary, we have described the basic features of matching, presented a logistic regression model for the analysis of matched data, and have illustrated the model using an example from a 2-to-1 matched case-control study. We have also discussed how to assess interaction of the matching variables with exposure, the issue of pooling exchangeable matched sets, and how to analyze matched follow-up data.

Logistic Regression Chapters

1. Introduction
2. Important Special Cases

⋮

✓ 11. Analysis of Matched Data

12. Polytomous Logistic Regression
13. Ordinal Logistic Regression

The reader may wish to review the detailed summary and to try the practice exercises and the test that follow.

Up to this point we have considered dichotomous outcomes only. In the next two chapters, the standard logistic model is extended to handle outcomes with three or more categories.

Detailed Outline

I. Overview (page 392)

Focus:

- Basics of matching
- Model for matched data
- Control of confounding and interaction
- Examples

II. Basic features of matching (pages 392–394)

A. Study design procedure: Select referent group to be constrained so as to be comparable to index group on one or more factors:

   i. Case-control study (our focus): referent = controls, index = cases

   ii. Follow-up study: referent = unexposed, index = exposed

B. Category matching: If case-control study, find, for each case, one or more controls in the same combined set of categories of matching factors

C. Types of matching: 1-to-1, *R*-to-1, other

D. To match or not to match:

   i. Advantage: Can gain efficiency/precision

   ii. Disadvantages: Costly to find matches and might lose information discarding controls

   iii. Safest strategy: Match on strong risk factors expected to be confounders

   iv. Validity not a reason for matching: Can get valid answer even when not matching

III. Matched analyses using stratification (pages 394–397)

A. Strata are matched sets, e.g., if 4-to-1 matching, each stratum contains five observations

B. Special case: 1-to-1 matching: four possible forms of strata:

   i. Both case and control are exposed (*W* pairs)

   ii. Only case is exposed (*X* pairs)

   iii. Only control is exposed (*Y* pairs)

   iv. Neither case nor control is exposed (*Z* pairs)

C. Two equivalent analysis procedures for 1-to-1 matching:

   i. Mantel–Haenszel (MH): Use MH test on all strata and compute MOR estimate of OR

   ii. McNemar approach: Group data by pairs (*W*, *X*, *Y*, and *Z* as in B above). Use McNemar's chi-square statistic $(X - Y)^2/(X + Y)$ for test and X/Y for estimate of OR

D. R-to-1 matching: Use MH test statistic and MOR

IV. The logistic model for matched data (pages 397–400)

A. Advantage: Provides an efficient analysis when there are variables other than matching variables to control.

B. Model uses dummy variables in identifying different strata.

C. Model form:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \sum \gamma_{1i} V_{1i} + \sum \gamma_{2j} V_{2j} + E \sum \delta_k W_k,$$

where V_{1i} are dummy variables identifying matched strata, V_{2j} are potential confounders based on variables not involved in the matching, and W_k are effect modifiers (usually) based on variables not involved in the matching.

D. Odds ratio expression if E is coded as (0, 1):

$$\text{ROR} = \exp\left(\beta + \sum \delta_k W_k\right).$$

V. An application (pages 400–403)

A. Case-control study, 2-to-1 matching, $D = \text{MI}$ (0, 1), $E = \text{SMK}$ (0, 1),

four matching variables: AGE, RACE, SEX, HOSPITAL,

two variables not matched: SBP, ECG,

$n = 117$ (39 matched sets, 3 observations per set).

B. Model form:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta\text{SMK} + \sum_{i=1}^{38} \gamma_{1i} V_{1i} + \gamma_{21}\text{SBP} + \gamma_{22}\text{ECG} + \text{SMK}(\delta_1\text{SBP} + \delta_2\text{ECG}).$$

C. Odds ratio:

$$\text{ROR} = \exp(\beta + \delta_1\text{SBP} + \delta_2\text{ECG}).$$

D. Analysis: Use conditional ML estimation; interaction not significant

No interaction model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta\text{SMK} + \sum_{i=1}^{38} \gamma_{1i} V_{1i} + \gamma_{21}\text{SBP} + \gamma_{22}\text{ECG}.$$

Odds ratio formula:

$$\text{ROR} = \exp(\beta),$$

Gold standard OR estimate controlling for SBP and ECG: 2.07, Narrowest CI obtained when only ECG is controlled: OR estimate is 2.08, Overall conclusion: OR approximately 2, but not significant.

VI. Assessing Interaction Involving Matching Variables (pages 404–406)

A. Option 1: Add product terms of the form $E \times V_{1i}$, where V_{1i} are dummy variables for matching strata.

$$\text{Model: logit P}(\mathbf{X}) = \alpha + \beta E + \sum \gamma_{1i} V_{1i} + \sum \gamma_{2j} V_{2j} + E \sum \delta_{1i} V_{1i} + E \sum \delta_k W_k,$$

where V_{2j} are other covariates (not matched) and W_k are effect modifiers defined from other covariates.

Criticism of option 1:

- Difficult to identify specific effect modifiers
- Number of parameters may exceed n

B. Option 2: Add product terms of the form $E \times W_{1\text{m}}$, where $W_{1\text{m}}$ are the matching variables in original form.

$$\text{Model: logit P}(\mathbf{X}) = \alpha + \beta E + \sum \gamma_{1i} V_{1i} + \sum \gamma_{2j} V_{2j} + E \sum \delta_{1i} W_{1\text{m}} + E \sum \delta_k W_k,$$

where V_{2j} are other covariates (not matched) and W_k are effect modifiers defined from other covariates.

Criticism of option 2:

- Model is not HWF (i.e., $E \times W_{1\text{m}}$ in model but not $W_{1\text{m}}$)

But, matching variables are in model in different ways as both effect modifiers and confounders.

C. Other alternatives:

- Do not match on any variable considered as an effect modifier
- Do not assess interaction for any matching variable

VII. Pooling Matching Strata (pages 407–409)

A. Example: Pair-match on SMK (0, 1), 100 cases, 60 matched pairs of smokers, 40 matched pairs of nonsmokers.

B. Controls for two or more matched pairs that have same SMK status are interchangeable.

Corresponding matched sets are called *exchangeable*.

C. Example (continued):

60 exchangeable smoker matched pairs.

40 exchangeable nonsmoker matched pairs.

D. Recommendation:

- Identify and pool exchangeable matched sets.
- Carry out stratified analysis or logistic regression using pooled strata.
- Consider using unconditional ML estimation (but conditional ML estimation always gives unbiased estimates).

E. Reason for pooling: Treating exchangeable matched sets as separate strata is artificial.

VIII. Analysis of Matched Follow-up Data (pages 409–413)

A. In follow-up studies, unexposed subjects are selected to have same distribution on matching variables as exposed subjects.

B. In follow-up studies, frequency matching rather than individual matching is typically used because of practical convenience and to obtain larger sample size.

C. Model same as for matched case-control studies except dummy variables defined by exposed/ unexposed matched sets:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \sum \gamma_{1i} V_{1i} + \sum \gamma_{2j} V_{2j} + E \sum \delta_k W_k.$$

D. Analysis if frequency matching used: Consider unconditional ML estimation when number of strata is small, although conditional ML estimation will always give unbiased answers.

E. Analysis if pair-matching is used and no pooling is done: Use McNemar approach that considers concordant and discordant pairs (P, Q, R, and S) and computes $\widehat{\text{MRR}} = (P+Q)/(P+R)$, $\widehat{\text{MOR}} = Q/R$, and $\chi^2_{\text{MH}} = (Q-R)^2/(Q+R)$.

F. Example: Pair-matched follow-up study with 4,830 matched pairs, E = VS (vasectomy status), D = MI (myocardial infarction status), match on AGE and YEAR (of follow-up); $P = 0$, $Q = 20$, $R = 16$, $S = 4790$.

$$\widehat{\text{MRR}} = 1.25 = \widehat{\text{MOR}},\ \chi^2_{\text{MH}} = 0.44(\text{N.S.}).$$

Criticisms:

- Information on 4,790 matched pairs not used
- Pooling exchangeable matched sets not used
- Frequency matching not used

G. Analysis that controls for both matched and unmatched variables: use logistic regression on only discordant pairs.

H. In matched follow-up studies, need only analyze discordant pairs. In matched case-control studies, use only discordant pairs, provided that there are no other control variables other than matching variables.

IX. Summary (page 414)

Practice Exercises

True or False (Circle T or F)

T F 1. In a case-control study, category pair-matching on age and sex is a procedure by which, for each control in the study, a case is found as its pair to be in the same age category and same sex category as the control.

T F 2. In a follow-up study, pair-matching on age is a procedure by which the age distribution of cases (i.e., those with the disease) in the study is constrained to be the same as the age distribution of noncases in the study.

T F 3. In a 3-to-1 matched case-control study, the number of observations in each stratum, assuming sufficient controls are found for each case, is four.

T F 4. An advantage of matching over not matching is that a more precise estimate of the odds ratio may be obtained from matching.

T F 5. One reason for deciding to match is to gain validity in estimating the odds ratio of interest.

T F 6. When in doubt, it is safer to match than not to match.

T F 7. A matched analysis can be carried out using a stratified analysis in which the strata consists of the collection of matched sets.

T F 8. In a pair-matched case-control study, the Mantel–Haenszel odds ratio (i.e., the MOR) is equivalent to McNemar's test statistic $(X - Y)^2/(X + Y)$. (Note: X denotes the number of pairs for which the case is exposed and the control is unexposed, and Y denotes the number of pairs for which the case is unexposed and the control is exposed.)

T F 9. When carrying out a Mantel–Haenszel chi-square test for 4-to-1 matched case-control data, the number of strata is equal to 5.

T F 10. Suppose in a pair-matched case-control study, that the number of pairs in each of the four cells of the table used for McNemar's test is given by $W = 50$, $X = 40$, $Y = 20$, and $Z = 100$. Then, the computed value of McNemar's test statistic is given by 2.

11. For the pair-matched case-control study described in Exercise 10, let E denote the (0, 1) exposure variable and let D denote the (0, 1) disease variable. State the

logit form of the logistic model that can be used to analyze these data. (Note: Other than the variables matched, there are no other control variables to be considered here.)

12. Consider again the pair-matched case-control data described in Exercise 10 ($W = 50$, $X = 40$, $Y = 20$, $Z = 100$). Using conditional ML estimation, a logistic model fitted to these data resulted in an estimated coefficient of exposure equal to 0.693, with standard error equal to 0.274. Using this information, compute an estimate of the odds ratio of interest and compare its value with the estimate obtained using the MOR formula X/Y.
13. For the same situation as in Exercise 12, compute the Wald test for the significance of the exposure variable and compare its squared value and test conclusion with that obtained using McNemar's test.
14. Use the information provided in Exercise 12 to compute a 95% confidence interval for the odds ratio, and interpret your result.
15. If unconditional ML estimation had been used instead of conditional ML estimation, what estimate would have been obtained for the odds ratio of interest? Which estimation method is correct, conditional or unconditional, for this data set?

Consider a 2-to-1 matched case-control study involving 300 bisexual males, 100 of whom are cases with positive HIV status, with the remaining 200 being HIV negative. The matching variables are AGE and RACE. Also, the following additional variables are to be controlled but are not involved in the matching: NP, the number of sexual partners within the past 3 years; ASCM, the average number of sexual contacts per month over the past 3 years, and PAR, a (0, 1) variable indicating whether or not any sexual partners in the past 5 years were in high-risk groups for HIV infection. The exposure variable is CON, a (0, 1) variable indicating whether the subject used consistent and correct condom use during the past 5 years.

16. Based on the above scenario, state the logit form of a logistic model for assessing the effect of CON on HIV acquisition, controlling for NP, ASCM, and PAR as potential confounders and PAR as the only effect modifier.
17. Using the model given in Exercise 16, give an expression for the odds ratio for the effect of CON on HIV status, controlling for the confounding effects of AGE,

RACE, NP, ASCM, and PAR, and for the interaction effect of PAR.

18. For the model used in Exercise 16, describe the strategy you would use to arrive at a final model that controls for confounding and interaction.

The data below are from a hypothetical pair-matched case-control study involving five matched pairs, where the only matching variable is smoking (SMK). The disease variable is called CASE and the exposure variable is called EXP. The matched set number is identified by the variable STRATUM.

ID	STRATUM	CASE	EXP	SMK
1	1	1	1	0
2	1	0	1	0
3	2	1	0	0
4	2	0	1	0
5	3	1	1	1
6	3	0	0	1
7	4	1	1	0
8	4	0	0	0
9	5	1	0	1
10	5	0	0	1

19. How many concordant pairs are there where both pair members are exposed?
20. How many concordant pairs are there where both members are unexposed?
21. How many discordant pairs are there where the case is exposed and the control is unexposed?
22. How many discordant pairs are there where case is unexposed and the control is exposed?

The table below summarizes the matched pairs information described in the previous questions.

		not *D*	
		E	not *E*
D	*E*	1	2
	not *E*	1	1

23. What is the estimated MOR for these data?
24. What type of matched analysis is being used with this table, pooled or unpooled? Explain briefly.

The table below groups the matched pairs information described in Exercises 19–22 into two smoking strata.

SMK = 1

	E	not *E*	
D	1	1	2
not *D*	0	2	2
			4

SMK = 0

	E	not *E*	
D	2	1	3
not *D*	2	1	3
			6

25. What is the estimated MOR from these data?
26. What type of matched analysis is being used here, pooled or unpooled?
27. Which type of analysis should be preferred for these matched data (where smoking status is the only matched variable), pooled or unpooled?

The data below switches the nonsmoker control of stratum 2 with the nonsmoker control of stratum 4 from the data set provided for Exercises 19–22. Let $W =$ no. of concordant ($E = 1$, $E = 1$) pairs, $X =$ no. of discordant ($E = 1$, $E = 0$) pairs, $Y =$ no. of discordant ($E = 0, E = 1$) pairs, and $Z =$ no. of concordant ($E = 0$, $E = 0$) pairs for the "switched" data.

ID	STRATUM	CASE	EXP	SMK
1	1	1	1	0
2	1	0	1	0
3	2	1	0	0
4	2	0	0	0
5	3	1	1	1
6	3	0	0	1
7	4	1	1	0
8	4	0	1	0
9	5	1	0	1
10	5	0	0	1

28. What are the values for *W, X, Y,* and *Z*?
29. What are the values of $\widehat{\text{MOR}}$ (unpooled) and $\widehat{\text{MOR}}$ (pooled)?

Based on the above data and your answers to the above Exercises:

30. Which of the following helps explain why the pooled $\widehat{\text{MOR}}$ should be preferred to the unpooled $\widehat{\text{MOR}}$? (Circle the best answer)
    a. The pooled $\widehat{\text{MOR}}$s are equal, whereas the unpooled $\widehat{\text{MOR}}$s are different.
    b. The unpooled $\widehat{\text{MOR}}$s assume that exchangeable matched pairs are not unique.
    c. The pooled $\widehat{\text{MOR}}$s assume that exchangeable matched pairs are unique.
    d. None of the choices a, b, and c above are correct.
    e. All of the choices a, b, and c above are correct.

Test

True or False (Circle T or F)

T F 1. In a category-matched 2-to-1 case-control study, each case is matched to two controls who are in the same category as the case for each of the matching factors.

T F 2. An advantage of matching over not matching is that information may be lost when not matching.

T F 3. If we do not match on an important risk factor for the disease, it is still possible to obtain an unbiased estimate of the odds ratio by doing an appropriate analysis that controls for the important risk factor.

T F 4. McNemar's test statistic is not appropriate when there is *R*-to-1 matching and *R* is at least 2.

T F 5. In a matched case-control study, logistic regression can be used when it is desired to control for variables involved in the matching as well as variables not involved in the matching.

6. Consider the following McNemar's table from the study analyzed by Donovan et al. (1984). This is a pair-matched case-control study, where the cases are babies born with genetic anomalies and controls are babies born without such anomalies. The matching variables are hospital, time period of birth, mother's age, and health insurance status. The exposure factor is status of father (Vietnam veteran = 1 or nonveteran = 0):

		Case	
		E	not *E*
Control	*E*	2	121
	not *E*	125	8254

For the above data, carry out McNemar's test for the significance of exposure and compute the estimated odds ratio. What are your conclusions?

7. State the logit form of the logistic model that can be used to analyze the study data.

8. The following printout results from using conditional ML estimation of an appropriate logistic model for analyzing the data:

					95% CI for OR	
Variable	β	s_β	*P*-value	OR	*L*	*U*
E	0.032	0.128	0.901	1.033	0.804	1.326

Use these results to compute the squared Wald test statistic for testing the significance of exposure and compare this test statistic with the McNemar chi-square statistic computed in Question 6.

9. How does the odds ratio obtained from the printout given in Question 8 compare with the odds ratio computed using McNemar's formula *X*/*Y*?
10. Explain how the confidence interval given in the printout is computed.

Answers to Practice Exercises

1. F: cases are selected first, and controls are matched to cases.
2. F: the age distribution for unexposed persons is constrained to be the same as for exposed persons.
3. T
4. T
5. F: matching is not needed to obtain a valid estimate of effect.
6. F: when in doubt, matching may not lead to increased precision; it is safe to match only if the potential matching factors are strong risk factors expected to be confounders in the data.
7. T
8. F: the Mantel–Haenszel chi-square statistic is equal to McNemar's test statistic.
9. F: the number of strata equals the number of matched sets.
10. F: the computed value of McNemar's test statistic is 6.67; the MOR is 2.
11. $\text{logit P}(\mathbf{X}) = \alpha + \beta E + \sum_{i=1}^{209} \gamma_{1i} V_{1i}$,

    where the V_{1i} denote dummy variables indicating the different matched pairs (strata).
12. Using the output, the estimated odds ratio is exp (0.693), which equals 1.9997. The $\widehat{\text{MOR}}$ is computed as *X*/*Y* equals 40/20 = 2. Thus, the estimate obtained using conditional logistic regression is equal to the $\widehat{\text{MOR}}$.
13. The Wald statistic, which is a *Z* statistic, is computed as 0.693/0.274, which equals 2.5292. This is significant at the 0.01 level of significance, i.e., *P* is less than 0.01. The squared Wald statistic, which has a chi-square distribution with one degree of freedom under the null hypothesis of no effect, is computed to be 6.40. The McNemar chi-square statistic is 6.67, which is quite similar to the Wald result, though not exactly the same.
14. The 95% confidence interval for the odds ratio is given by the formula $\exp\left[\hat{\beta} \pm 1.96\sqrt{\widehat{\text{var}}\left(\hat{\beta}\right)}\right]$,

    which is computed to be
    $\exp(0.693 \pm 1.96 \times 0.274) = \exp(0.693 \pm 0.53704)$,
    which equals $(e^{0.15596}, e^{1.23004}) = (1.17, 3.42)$.
    This confidence interval around the point estimate of 2 indicates that the point estimate is somewhat unstable. In particular, the lower limit is close to the null value of 1, whereas the upper limit is close to 4. Note also that the confidence interval does not include the

null value, which supports the statistical significance found in Exercise 13.

15. If unconditional ML estimation had been used, the odds ratio estimate would be higher (i.e., an overestimate) than the estimate obtained using conditional ML estimation. In particular, because the study involved pair-matching, the unconditional odds ratio is the square of the conditional odds ratio estimate. Thus, for this dataset, the conditional estimate is given by $\widehat{\text{MOR}}$ equal to 2, whereas the unconditional estimate is given by the square of 2 or 4. The correct estimate is 2, not 4.

16.

$$\text{logit P}(\mathbf{X}) = \alpha + \beta\text{CON} + \sum_{i=1}^{99} \gamma_{1i}V_{1i} + \gamma_{21}\text{NP} + \gamma_{22}\text{ASCM}$$
$$+ \gamma_{23}\text{PAR} + \delta\text{CON} \times \text{PAR},$$

where the V_{1i} are 99 dummy variables indicating the 100 matching strata, with each stratum containing three observations.

17. $\widehat{\text{ROR}} = \exp\left(\hat{\beta} + \hat{\delta}\text{PAR}\right)$.

18. A recommended strategy for model building involves first testing for the significance of the interaction term in the starting model given in Exercise 16. If this test is significant, then the final model must contain the interaction term, the main effect of PAR (from the Hierarchy Principle), and the 99 dummy variables for matching. The other two variables NP and ASCM may be dropped as nonconfounders if the odds ratio given by Exercise 17 does not meaningfully change when either or both variables are removed from the model. If the interaction test is not significant, then the reduced (no interaction) model is given by the expression

$$\text{logit P}(\mathbf{X}) = \alpha + \beta\text{CON} + \sum_{i=1}^{99} \gamma_{1i}V_{1i} + \gamma_{21}\text{NP}$$
$$+ \gamma_{22}\text{ASCM} + \gamma_{23}\text{PAR}.$$

Using this reduced model, the odds ratio formula is given by $\exp(\beta)$, where β is the coefficient of the CON variable. The final model must contain the 99 dummy variables which incorporate the matching into the model. However, NP, ASCM, and/or PAR may be dropped as nonconfounders if the odds ratio $\exp(\beta)$ does not change when one or more of these three variables are dropped from the model. Finally, precision of the estimate needs to be considered by comparing confidence intervals for the odds ratio. If a meaningful gain of precision is made by dropping a nonconfounder, then such a nonconfounder may be dropped. Otherwise (i.e., no gain in precision), the nonconfounder should remain in the model with all other variables needed for controlling confounding.

19. 1
20. 1
21. 2
22. 1
23. 2
24. Unpooled; the analysis treats all five strata (matched pairs) as unique.
25. 2.5
26. Pooled.
27. Pooled; treating the five strata as unique is artificial since there are exchangeable strata that should be pooled.
28. $W = 1, X = 1, Y = 0$, and $Z = 2$.
29. mOR(unpooled) = undefined; mOR(pooled) = 2.5.
30. Only choice a is correct.

12 Polytomous Logistic Regression

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_12, © Springer Science+Business Media, LLC 2010

Introduction

In this chapter, the standard logistic model is extended to handle outcome variables that have more than two categories. Polytomous logistic regression is used when the categories of the outcome variable are nominal, that is, they do not have any natural order. When the categories of the outcome variable do have a natural order, ordinal logistic regression may also be appropriate.

The focus of this chapter is on polytomous logistic regression. The mathematical form of the polytomous model and its interpretation are developed. The formulas for the odds ratio and confidence intervals are derived, and techniques for testing hypotheses and assessing the statistical significance of independent variables are shown.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

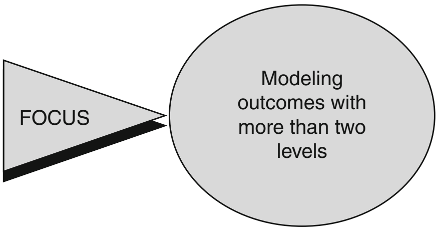
Objectives

Upon completing this chapter, the learner should be able to:

1. State or recognize the difference between nominal and ordinal variables.
2. State or recognize when the use of polytomous logistic regression may be appropriate.
3. State or recognize the polytomous regression model.
4. Given a printout of the results of a polytomous logistic regression:
   a. State the formula and compute the odds ratio
   b. State the formula and compute a confidence interval for the odds ratio
   c. Test hypotheses about the model parameters using the likelihood ratio test or the Wald test, stating the null hypothesis and the distribution of the test statistic with the corresponding degrees of freedom under the null hypothesis
5. Recognize how running a polytomous logistic regression differs from running multiple standard logistic regressions.

Presentation

I. Overview



This presentation and the presentation that follows describe approaches for extending the standard logistic regression model to accommodate a disease, or outcome, variable that has more than two categories. Up to this point, our focus has been on models that involve a dichotomous outcome variable, such as disease present/absent. However, there may be situations in which the investigator has collected data on multiple levels of a single outcome. We describe the *form* and key *characteristics* of one model for such multilevel outcome variables: the polytomous logistic regression model.

Examples of multilevel outcomes:

1. Absent, mild, moderate, severe
2. In situ, locally invasive, metastatic
3. Choice of treatment regimen

Examples of outcome variables with more than two levels might include (1) disease symptoms that have been classified by subjects as being absent, mild, moderate, or severe, (2) invasiveness of a tumor classified as in situ, locally invasive, or metastatic, or (3) patients' preferred treatment regimen, selected from among three or more options.

One approach: dichotomize outcome

Change

0	1	2

to

0 1	2

One possible approach to the analysis of data with a polytomous outcome would be to choose an appropriate cut-point, dichotomize the multilevel outcome variable, and then simply utilize the logistic modeling techniques discussed in previous chapters.

EXAMPLE

Change

None	Mild	Moderate	Severe

to

None or mild	Moderate or severe

For example, if the outcome symptom severity has four categories of severity, one might compare subjects with none or only mild symptoms to those with either moderate or severe symptoms.

Disadvantage of dichotomizing: Loss of detail (e.g., mild vs. none? moderate vs. mild?)

The disadvantage of dichotomizing a polytomous outcome is loss of detail in describing the outcome of interest. For example, in the scenario given above, we can no longer compare mild vs. none or moderate vs. mild. This loss of detail may, in turn, affect the conclusions made about the exposure–disease relationship.

Alternate approach: Use model for a polytomous outcome

Nominal or ordinal outcome?

The detail of the original data coding can be retained through the use of models developed specifically for polytomous outcomes. The specific form that the model takes depends, in part, on whether the multilevel outcome variable is measured on a nominal or an ordinal scale.

Nominal: Different categories; no ordering

EXAMPLE

Endometrial cancer subtypes:

- Adenosquamous
- Adenocarcinoma
- Other

Nominal variables simply indicate different categories. An example is histological subtypes of cancer. For endometrial cancer, three possible subtypes are adenosquamous, adenocarcinoma, and other.

Ordinal: Levels have natural ordering

EXAMPLE

Tumor grade:

- Well differentiated
- Moderately differentiated
- Poorly differentiated

Ordinal variables have a natural ordering among the levels. An example is cancer tumor grade, ranging from well differentiated to moderately differentiated to poorly differentiated tumors.

Nominal outcome ⇒ Polytomous model

Ordinal outcome ⇒ Ordinal model or polytomous model

An outcome variable that has three or more nominal categories can be modeled using polytomous logistic regression. An outcome variable with three or more ordered categories can also be modeled using polytomous regression, but can also be modeled with ordinal logistic regression, provided that certain assumptions are met. Ordinal logistic regression is discussed in detail in Chap. 13.

II. Polytomous Logistic Regression: An Example with Three Categories

$$E \xRightarrow{?} D$$

When modeling a multilevel outcome variable, the epidemiological question remains the same: What is the relationship of one or more exposure or study variables (E) to a disease or illness outcome (D)?

EXAMPLE

Simplest case of polytomous model:

- Outcome with three categories
- One dichotomous exposure variable

In this section, we present an example of a polytomous logistic regression model with one dichotomous exposure variable and an outcome (D) that has three categories. This is the simplest case of a polytomous model. Later in the presentation, we discuss extending the polytomous model to more than one predictor variable and then to outcomes with more than three categories.

Data source:
Black/White Cancer Survival Study

$$E = \text{AGEGP}\begin{cases} 0 & \text{if } 50\text{–}64 \\ 1 & \text{if } 65\text{–}79 \end{cases}$$

$$D = \text{SUBTYPE}\begin{cases} 0 & \text{if Adenocarcinoma} \\ 1 & \text{if Adenosquamous} \\ 2 & \text{if Other} \end{cases}$$

The example uses data from the National Cancer Institute's Black/White Cancer Survival Study (Hill et al., 1995). Suppose we are interested in assessing the effect of age group on histological subtype among women with primary endometrial cancer. AGEGP, the exposure variable, is coded as 0 for aged 50–64 or 1 for aged 65–79. The disease variable, histological subtype, is coded 0 for adenocarcinoma, 1 for adenosquamous, and 2 for other.

SUBTYPE (0, 1, 2) uses arbitrary coding.

There is no inherent order in the outcome variable. The 0, 1, and 2 coding of the disease categories is arbitrary.

	AGEGP	
	50–64 $E = 0$	65–79 $E = 1$
Adenocarcinoma $D = 0$	77	109
Adenosquamous $D = 1$	11	34
Other $D = 2$	18	39

The 3 × 2 table of the data is presented on the left.

Outcome categories:

A B C D

Reference (arbitrary choice)

Then compare:

A vs. C, B vs. C, and D vs. C

With polytomous logistic regression, one of the categories of the outcome variable is designated as the reference category and each of the other levels is compared with this reference. The choice of reference category can be arbitrary and is at the discretion of the researcher. See example at left. Changing the reference category does not change the form of the model, but it does change the interpretation of the parameter estimates in the model.

EXAMPLE (*continued*)

Reference group = Adenocarcinoma

Two comparisons:

1. Adenosquamous ($D = 1$) vs. Adenocarcinoma ($D = 0$)
2. Other ($D = 2$) vs. Adenocarcinoma ($D = 0$)

Using data from table:

$$\widehat{\text{OR}}_{1\text{ vs. }0} = \frac{77 \times 34}{109 \times 11} = 2.18$$

$$\widehat{\text{OR}}_{2\text{ vs. }0} = \frac{77 \times 39}{109 \times 18} = 1.53$$

In our three-outcome example, the Adenocarcinoma group has been designated as the reference category. We are therefore interested in modeling two main comparisons. We want to compare subjects with an Adenosquamous outcome (category 1) to those subjects with an Adenocarcinoma outcome (category 0) and we also want to compare subjects with an Other outcome (category 2) to those subjects with an Adenocarcinoma outcome (category 0).

If we consider these two comparisons separately, the crude odds ratios can be calculated using data from the preceding table. The crude odds ratio comparing Adenosquamous (category 1) to Adenocarcinoma (category 0) is the product of 77 and 34 divided by the product of 109 and 11, which equals 2.18. Similarly, the crude odds ratio comparing Other (category 2) to Adenocarcinoma (category 0) is the product of 77 and 39 divided by the product of 109 and 18, which equals 1.53.

Dichotomous vs. polytomous model: Odds vs. "odds-like" expressions

$$\text{logit P}(\mathbf{X}) = \ln\left[\frac{\text{P}(D = 1 \mid \mathbf{X})}{\text{P}(D = 0 \mid \mathbf{X})}\right] = \alpha + \sum_{i=1}^{k} \beta_i X_i$$

Recall that for a dichotomous outcome variable coded as 0 or 1, the logit form of the logistic model, logit P(**X**), is defined as the natural log of the odds for developing a disease for a person with a set of independent variables specified by **X**. This logit form can be written as the linear function shown on the left.

Odds of disease: a ratio of probabilities

Dichotomous outcome:

$$\text{odds} = \frac{\mathrm{P}(D=1)}{1-\mathrm{P}(D=1)} = \frac{\mathrm{P}(D=1)}{\mathrm{P}(D=0)}$$

The odds for developing disease can be viewed as a ratio of probabilities. For a dichotomous outcome variable coded 0 and 1, the odds of disease equal the probability that disease equals 1 divided by 1 minus the probability that disease equals 1, or the probability that disease equals 1 divided by the probability that disease equals 0.

Polytomous outcome (three categories):

Use "odds-like" expressions for two comparisons

$$(1)\ \frac{\mathrm{P}(D=1)}{\mathrm{P}(D=0)} \qquad (2)\ \frac{\mathrm{P}(D=2)}{\mathrm{P}(D=0)}$$

For polytomous logistic regression with a three-level variable coded 0, 1, and 2, there are two analogous expressions, one for each of the two comparisons we are making. These expressions are also in the form of a ratio of probabilities.

The logit form of model uses ln of "odds-like" expressions

$$(1)\ \ln\left[\frac{\mathrm{P}(D=1)}{\mathrm{P}(D=0)}\right] \ (2)\ \ln\left[\frac{\mathrm{P}(D=2)}{\mathrm{P}(D=0)}\right]$$

In polytomous logistic regression with three levels, we therefore define our model using two expressions for the natural log of these "odds-like" quantities. The first is the natural log of the probability that the outcome is in category 1 divided by the probability that the outcome is in category 0; the second is the natural log of the probability that the outcome is in category 2 divided by the probability that the outcome is in category 0.

$$\mathrm{P}(D=0)+\mathrm{P}(D=1)+\mathrm{P}(D=2)=1$$

BUT

$$\mathrm{P}(D=1)+\mathrm{P}(D=0)\neq 1$$

$$\mathrm{P}(D=2)+\mathrm{P}(D=0)\neq 1$$

Therefore:

$$\frac{\mathrm{P}(D=1)}{\mathrm{P}(D=0)} \text{ and } \frac{\mathrm{P}(D=2)}{\mathrm{P}(D=0)}$$

"odds-like" but not true odds (unless analysis restricted to two categories)

When there are three categories of the outcome, the sum of the probabilities for the three outcome categories must be equal to 1, the total probability. Because each comparison considers only two probabilities, the probabilities in the ratio do not sum to 1. Thus, the two "odds-like" expressions are not true odds. However, if we restrict our interest to just the two categories being considered in a given ratio, we may still conceptualize the expression as an odds. In other words, each expression is an odds *only* if we condition on the outcome being in one of the two categories of interest. For ease of the subsequent discussion, we will use the term "odds" rather than "odds-like" for these expressions.

Model for three categories, one predictor (X_1 = AGEGP):

$$\ln\left[\frac{P(D=1\,|\,X_1)}{P(D=0\,|\,X_1)}\right]=\alpha_1+\beta_{11}X_1$$

Because our example has three outcome categories and one predictor (i.e., AGEGP), our polytomous model requires two regression expressions. One expression gives the log of the probability that the outcome is in category 1 divided by the probability that the outcome is in category 0, which equals α_1 plus β_{11} times X_1.

$$\ln\left[\frac{P(D=2\,|\,X_1)}{P(D=0\,|\,X_1)}\right]=\alpha_2+\beta_{21}X_1$$

We are also *simultaneously* modeling the log of the probability that the outcome is in category 2 divided by the probability that the outcome is in category 0, which equals α_2 plus β_{21} times X_1.

$$\begin{tikzcd}
\alpha_1 \arrow[dr] & & \beta_{11} \arrow[dl] \\
& 1 \text{ vs. } 0 & \\
\alpha_2 \arrow[dr] & & \beta_{21} \arrow[dl] \\
& 2 \text{ vs. } 0 &
\end{tikzcd}$$

Both the alpha and beta terms have a subscript to indicate which comparison is being made (i.e., category **1** vs. 0 or category **2** vs. 0).

III. Odds Ratio with Three Categories

$$\left.\begin{array}{cc}\hat{\alpha}_1 & \hat{\alpha}_2 \\ \hat{\beta}_{11} & \hat{\beta}_{21}\end{array}\right\}\ \begin{array}{l}\text{Estimates obtained}\\ \text{as in SLR}\end{array}$$

Once a polytomous logistic regression model has been fit and the parameters (intercepts and beta coefficients) have been estimated, we can then calculate estimates of the disease–exposure association in a similar manner to the methods used in standard logistic regression (SLR).

Special case for one predictor where $X_1 = 1$ or $X_1 = 0$

Consider the special case in which the only independent variable is the exposure variable and the exposure is coded 0 and 1. To assess the effect of the exposure on the outcome, we compare $X_1 = 1$ to $X_1 = 0$.

Two odds ratios:

OR_1 (category 1 vs. category 0) (Adenosquamous vs. Adenocarcinoma)

OR_2 (category 2 vs. category 0) (Other vs. Adenocarcinoma)

We need to calculate two odds ratios, one that compares category 1 (Adenosquamous) to category 0 (Adenocarcinoma) and one that compares category 2 (Other) to category 0 (Adenocarcinoma).

Recall that we are actually calculating a ratio of two "odds-like" expressions. However, we continue the conventional use of the term odds ratio for our discussion.

$$OR_1 = \frac{[P(D=1|X=1)/P(D=0|X=1)]}{[P(D=1|X=0)/P(D=0|X=0)]}$$

$$OR_2 = \frac{[P(D=2|X=1)/P(D=0|X=1)]}{[P(D=2|X=0)/P(D=0|X=0)]}$$

Each odds ratio is calculated in a manner similar to that used in standard logistic regression. The two OR formulas are shown on the left.

Adenosquamous vs. Adenocarcinoma:

$$OR_1 = \frac{\exp[\alpha_1 + \beta_{11}(1)]}{\exp[\alpha_1 + \beta_{11}(0)]} = e^{\beta_{11}}$$

Other vs. Adenocarcinoma:

$$OR_2 = \frac{\exp[\alpha_2 + \beta_{21}(1)]}{\exp[\alpha_2 + \beta_{21}(0)]} = e^{\beta_{21}}$$

Using our previously defined probabilities of the log odds, we substitute the two values of X_1 for the exposure (i.e., 0 and 1) into those expressions. After dividing, we see that the odds ratio for the first comparison (Adenosquamous vs. Adenocarcinoma) is e to the β_{11}.

The odds ratio for the second comparison (Other vs. Adenocarcinoma) is e to the β_{21}.

$OR_1 = e^{\beta_{11}}$ $\quad$ $OR_2 = e^{\beta_{21}}$

They are different!

We obtain two different odds ratio expressions, one utilizing β_{11} and the other utilizing β_{21}. Thus, quantifying the association between the exposure and outcome depends on which levels of the outcome are being compared.

General case for one predictor

$$\mathrm{OR}_g = \exp\left[\beta_{g1}\left(X_1^{**} - X_1^{*}\right)\right], \text{ where } g = 1, 2$$

The special case of a dichotomous predictor can be generalized to include categorical or continuous predictors. To compare any two levels ($X_1 = X_1^{**}$ vs. $X_1 = X_1^{*}$) of a predictor, the odds ratio formula is e to the β_{g1} times $(X_1^{**} - X_1^{*})$, where g defines the category of the disease variable (1 or 2) being compared with the reference category (0).

Computer output for polytomous model:

Is output listed in ascending or descending order?

EXAMPLE

SAS

Reference category: $D = 0$
Parameters for $D = 2$ comparison precede $D = 1$ comparison.

Variable	Estimate symbol
Intercept 1	$\hat{\alpha}_2$
Intercept 2	$\hat{\alpha}_1$
X_1	$\hat{\beta}_{21}$
X_1	$\hat{\beta}_{11}$

The output generated by a computer package for polytomous logistic regression includes alphas and betas for the log odds terms being modeled. Packages vary in the presentation of output, and the coding of the variables must be considered to correctly read and interpret the computer output for a given package. For example, in SAS, if $D = 0$ is designated as the reference category, the output is listed in descending order (see Appendix). This means that the listing of parameters pertaining to the comparison with category $D = 2$ precedes the listing of parameters pertaining to the comparison with category $D = 1$, as shown on the left.

EXAMPLE

Variable	Estimate	S.E.	Symbol
Intercept 1	−1.4534	0.2618	$\hat{\alpha}_2$
Intercept 2	−1.9459	0.3223	$\hat{\alpha}_1$
AGEGP	0.4256	0.3215	$\hat{\beta}_{21}$
AGEGP	0.7809	0.3775	$\hat{\beta}_{11}$

The results for the polytomous model examining histological subtype and age are presented on the left. The results were obtained from running PROC LOGISTIC in SAS. See the Computer Appendix for computer coding.

There are two sets of parameter estimates. The output is listed in descending order, with α_2 labeled as Intercept 1 and α_1 labeled as intercept 2. If $D = 2$ had been designated as the reference category, the output would have been in ascending order.

EXAMPLE (*continued*)

Other vs. Adenocarcinoma:

$$\ln\left[\frac{\hat{\mathrm{P}}(D=2\,|\,X_1)}{\hat{\mathrm{P}}(D=0\,|\,X_1)}\right]=-1.4534+(0.4256)\mathrm{AGEGP}$$

$$\widehat{\mathrm{OR}}_2=\exp[\hat{\beta}_{21}]=\exp(0.4256)=1.53$$

Adenosquamous vs. Adenocarcinoma:

$$\ln\left[\frac{\hat{\mathrm{P}}(D=1\,|\,X_1)}{\hat{\mathrm{P}}(D=0\,|\,X_1)}\right]=-1.9459+(0.7809)\mathrm{AGEGP}$$

$$\mathrm{OR}_1=\exp[\hat{\beta}_{11}]=\exp(0.7809)=2.18$$

Special case

One dichotomous exposure ⇒ polytomous model ORs = crude ORs

Interpretation of ORs

For older vs. younger subjects:

- Other tumor category more likely than Adenocarcinoma ($\widehat{\mathrm{OR}}_2 = 1.53$)
- Adenosquamous even more likely than Adenocarcinoma ($\widehat{\mathrm{OR}}_1 = 2.18$)

The equation for the estimated log odds of Other (category 2) vs. Adenocarcinoma (category 0) is negative 1.4534 plus 0.4256 times age group.

Exponentiating the beta estimate for age in this model yields an estimated odds ratio of 1.53.

The equation for the estimated log odds of Adenosquamous (category 1) vs. Adenocarcinoma (category 0) is negative 1.9459 plus 0.7809 times age group.

Exponentiating the beta estimate for AGEGP in this model yields an estimated odds ratio of 2.18.

The odds ratios from the polytomous model (i.e., 1.53 and 2.18) are the same as those we obtained earlier when calculating the crude odds ratios from the data table before modeling. In the special case, where there is one *dichotomous* exposure variable, the crude estimate of the odds ratio will match the estimate of the odds ratio obtained from a polytomous model (or from a standard logistic regression model).

We can interpret the odds ratios by saying that, for women diagnosed with primary endometrial cancer, older subjects (aged 65–79) relative to younger subjects (aged 50–64) were more likely to have their tumors categorized as Other than as Adenocarcinoma ($\widehat{\mathrm{OR}}_2 = 1.53$) and were even more likely to have their tumors classified as Adenosquamous than as Adenocarcinoma ($\widehat{\mathrm{OR}}_1 = 2.18$).

Interpretation of alphas

Log odds where all Xs set to 0.

Not informative if sampling done by outcome (i.e., "disease") status.

What is the interpretation of the alpha coefficients? They represent the log of the odds where all independent variables are set to zero (i.e., $X_i = 0$ for $i = 1$ to k). The intercepts are not informative, however, if sampling is done by outcome (i.e., disease status). For example, suppose the subjects in the endometrial cancer example had been selected based on tumor type, with age group (i.e., exposure status) determined after selection. This would be analogous to a case-control study design. Although the intercepts are not informative in this setting, the odds ratio is still a valid measure with this sampling method.

IV. Statistical Inference with Three Categories

Two types of inferences:

1. Hypothesis testing about parameters
2. Interval estimation around parameters

Procedures for polytomous outcomes or generalizations of SLR

In polytomous logistic regression, as with standard logistic regression (i.e., a dichotomous outcome), two types of statistical inferences are often of interest: (1) testing hypotheses and (2) deriving interval estimates around parameters. Procedures for both of these are straightforward generalizations of those that apply to logistic regression modeling with a dichotomous outcome variable (i.e., SLR).

95% CI for OR (one predictor)

$$\exp\left\{\hat{\beta}_{g1}\left(X_1^{**} - X_1^{*}\right) \pm 1.96\left(X_1^{**} - X_1^{*}\right)s_{\hat{\beta}_{g1}}\right\}$$

The confidence interval estimation is analogous to the standard logistic regression situation. For one predictor variable, with any levels (X_1^{**} and X_1^{*}) of that variable, the large-sample formula for a 95% confidence interval is of the general form shown at left.

EXAMPLE

Estimated standard errors:
(X_1 = AGEGP)

$s_{\hat{\beta}_{21}} = 0.3215, \quad s_{\hat{\beta}_{11}} = 0.3775$

Continuing with the endometrial cancer example, the estimated standard errors for the parameter estimates for AGEGP are 0.3215 for $\hat{\beta}_{21}$ and 0.3775 for $\hat{\beta}_{11}$.

EXAMPLE (*continued*)

95% CI for OR_2

$$= \exp[0.4256 \pm 1.96(0.3215)]$$
$$= (0.82, 2.87)$$

95% CI for OR_1

$$= \exp[0.7809 \pm 1.96(0.3775)]$$
$$= (1.04, 4.58)$$

The 95% confidence interval for OR_2 is calculated as 0.82 to 2.87, as shown on the left. The 95% confidence interval for OR_1 is calculated as 1.04 to 4.58.

Likelihood ratio test

Assess significance of X_1

2 βs tested at the same time

$$\Downarrow$$

2 degrees of freedom

As with a standard logistic regression, we can use a likelihood ratio test to assess the significance of the independent variable in our model. We must keep in mind, however, that rather than testing one beta coefficient for an independent variable, we are now testing two at the same time. There is a coefficient for each comparison being made (i.e., $D = 2$ vs. $D = 0$ and $D = 1$ vs. $D = 0$). This affects the number of parameters tested and, therefore, the degrees of freedom associated with the test.

EXAMPLE

3 levels of D and 1 predictor

$$\Downarrow$$

2 αs and 2 βs

Full model:

$$\ln\left[\frac{\text{P}(D = g \mid X_1)}{\text{P}(D = 0 \mid X_1)}\right] = \alpha_g + \beta_{g1}X_1, \quad g = 1, 2$$

Reduced model:

$$\ln\left[\frac{\text{P}(D = g)}{\text{P}(D = 0)}\right] = \alpha_g, \quad g = 1, 2$$

H_0: $\beta_{11} = \beta_{21} = 0$

Likelihood ratio test statistic:

$$-2\ln L_{\text{reduced}} - (-2\ln L_{\text{full}}) \sim \chi^2$$

with df = number of parameters set to zero under H_0

In our example, we have a three-level outcome variable and a single predictor variable, the exposure. As the model indicates, we have two intercepts and two beta coefficients.

If we are interested in testing for the significance of the beta coefficient corresponding to the exposure, we begin by fitting a full model (with the exposure variable in it) and then comparing that to a reduced model containing only the intercepts.

The null hypothesis is that the beta coefficients corresponding to the exposure variable are both equal to zero.

The likelihood ratio test is calculated as negative two times the log likelihood ($\ln L$) from the reduced model minus negative two times the log likelihood from the full model. The resulting statistic is distributed approximately chi-square, with degrees of freedom (df) equal to the number of parameters set equal to zero under the null hypothesis.

EXAMPLE

	$-2\ln L$
Reduced:	514.4
Full:	508.9

Difference = 5.5
df = 2
P-value = 0.06

In the endometrial cancer example, negative two times the log likelihood for the reduced model is 514.4, and for the full model is 508.9. The difference is 5.5. The chi-square P-value for this test statistic, with two degrees of freedom, is 0.06. The two degrees of freedom are for the two beta coefficients being tested, one for each comparison. We conclude that AGEGP is statistically significant at the 0.10 level but not at the 0.05 level.

Wald test

β for single outcome level tested

Whereas the likelihood ratio test allows for the assessment of the effect of an independent variable across all levels of the outcome simultaneously, it is possible that one might be interested in evaluating the effect of the independent variable at a single outcome level. A Wald test can be performed in this situation.

For two levels:

H_0: $\beta_{11} = 0 \quad H_0$: $\beta_{21} = 0$

$$Z = \frac{\hat{\beta}_{g1}}{s_{\hat{\beta}_{g1}}} \sim N(0,1)$$

The null hypothesis, for each level of interest, is that the beta coefficient is equal to zero. The Wald test statistics are computed as described earlier, by dividing the estimated coefficient by its standard error. This test statistic has an approximate normal distribution.

EXAMPLE

H_0: $\beta_{11} = 0$ (category 1 vs. 0)

$$Z = \frac{0.7809}{0.3775} = 2.07, \quad P = 0.04$$

H_0: $\beta_{21} = 0$ (category 2 vs. 0)

$$Z = \frac{0.4256}{0.3215} = 1.32, \quad P = 0.19$$

Continuing with our example, the null hypothesis for the Adenosquamous vs. Adenocarcinoma comparison (i.e., category 1 vs. 0) is that β_{11} equals zero. The Wald statistic for β_{11} is equal to 2.07, with a P-value of 0.04. The null hypothesis for the Other vs. Adenocarcinoma comparison (i.e., category 2 vs. 0) is that β_{21} equals zero. The Wald statistic for β_{21} is equal to 1.32, with a P-value of 0.19.

Conclusion: Is AGEGP significant?

- ⇒ Yes: Adenocarcinoma vs. Adenosquamous
- ⇒ No: Other vs. Adenosquamous.

At the 0.05 level of significance, we reject the null hypothesis for β_{11} but not for β_{21}. We conclude that AGEGP is statistically significant for the Adenosquamous vs. Adenocarcinoma comparison (category 1 vs. 0), but not for the Other vs. Adenocarcinoma comparison (category 2 vs. 0).

Decision: Retain or drop *both* β_{11} and β_{21} from model

We must either keep both betas (β_{11} and β_{21}) for an independent variable or drop both betas when modeling in polytomous regression. Even if only one beta is significant, both betas must be retained if the independent variable is to remain in the model.

V. Extending the Polytomous Model to *G* Outcomes and *k* Predictors

Adding more independent variables

Expanding the model to add more independent variables is straightforward. We can add k independent variables for each of the outcome comparisons.

$$\ln\left[\frac{\mathrm{P}(D=1\,|\,\mathbf{X})}{\mathrm{P}(D=0\,|\,\mathbf{X})}\right] = \alpha_1 + \sum_{i=1}^{k} \beta_{1i}X_i$$

$$\ln\left[\frac{\mathrm{P}(D=2\,|\,\mathbf{X})}{\mathrm{P}(D=0\,|\,\mathbf{X})}\right] = \alpha_2 + \sum_{i=1}^{k} \beta_{2i}X_i$$

The log odds comparing category 1 to category 0 is equal to α_1 plus the summation of the k independent variables times their β_1 coefficients. The log odds comparing category 2 to category 0 is equal to α_2 plus the summation of the k independent variables times their β_2 coefficients.

Same procedures for OR, CI, and hypothesis testing

The procedures for calculation of the odds ratios, confidence intervals, and for hypothesis testing remain the same.

EXAMPLE

$$D = \text{SUBTYPE}\begin{cases} 0 & \text{if Adenocarcinoma} \\ 1 & \text{if Adenosquamous} \\ 2 & \text{if Other} \end{cases}$$

Predictors

$X_1 = \text{AGEGP}$
$X_2 = \text{ESTROGEN}$
$X_3 = \text{SMOKING}$

To illustrate, we return to our endometrial cancer example. Suppose we wish to consider the effects of estrogen use and smoking status as well as AGEGP on histological subtype ($D = 0, 1, 2$). The model now contains three predictor variables: $X_1 = \text{AGEGP}$, $X_2 = \text{ESTROGEN}$, and $X_3 = \text{SMOKING}$.

EXAMPLE (*continued*)

$$X_1 = \text{AGEGP}\begin{cases} 0 & \text{if } 50\text{–}64 \\ 1 & \text{if } 65\text{–}79 \end{cases}$$

$$X_2 = \text{ESTROGEN}\begin{cases} 0 & \text{if never user} \\ 1 & \text{if ever user} \end{cases}$$

$$X_3 = \text{SMOKING}\begin{cases} 0 & \text{if former or never smoker} \\ 1 & \text{if current smoker} \end{cases}$$

Recall that AGEGP is coded as 0 for aged 50–64 or 1 for aged 65–79. Both estrogen use and smoking status are also coded as dichotomous variables. ESTROGEN is coded as 1 for ever user and 0 for never user. SMOKING is coded as 1 for current smoker and 0 for former or never smoker.

Adenosquamous vs. Adenocarcinoma:

$$\ln\left[\frac{\mathrm{P}(D=1\,|\,\mathbf{X})}{\mathrm{P}(D=0\,|\,\mathbf{X})}\right] = \alpha_1 + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3$$

The log odds comparing Adenosquamous ($D = 1$) to Adenocarcinoma ($D = 0$) is equal to α_1 plus β_{11} times X_1 plus β_{12} times X_2 plus β_{13} times X_3.

Other vs. Adenocarcinoma:

$$\ln\left[\frac{\mathrm{P}(D=2\,|\,\mathbf{X})}{\mathrm{P}(D=0\,|\,\mathbf{X})}\right] = \alpha_2 + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3$$

Similarly, the log odds comparing Other type ($D = 2$) to Adenocarcinoma ($D = 0$) is equal to α_2 plus β_{21} times X_1 plus β_{22} times X_2 plus β_{23} times X_3.

Variable	Estimate	S.E.	Symbol
Intercept 1	−1.2032	0.3190	$\hat{\alpha}_2$
Intercept 2	−1.8822	0.4025	$\hat{\alpha}_1$
AGEGP	0.2823	0.3280	$\hat{\beta}_{21}$
AGEGP	0.9871	0.4118	$\hat{\beta}_{11}$
ESTROGEN	−0.1071	0.3067	$\hat{\beta}_{22}$
ESTROGEN	−0.6439	0.3436	$\hat{\beta}_{12}$
SMOKING	−1.7913	1.0460	$\hat{\beta}_{23}$
SMOKING	0.8895	0.5254	$\hat{\beta}_{13}$

The output for the analysis is shown on the left. There are two beta estimates for each of the three predictor variables in the model. Thus, there are a total of eight parameters in the model, including the intercepts.

EXAMPLE (*continued*)

Adenosquamous vs. Adenocarcinoma:

$$\widehat{\mathrm{OR}}_1 = \frac{\exp[\hat{\alpha}_1 + \hat{\beta}_{11}(1) + \hat{\beta}_{12}(X_2) + \hat{\beta}_{13}(X_3)]}{\exp[\hat{\alpha}_1 + \hat{\beta}_{11}(0) + \hat{\beta}_{12}(X_2) + \hat{\beta}_{13}(X_3)]}$$
$$= \exp\hat{\beta}_{11} = \exp(0.9871) = 2.68$$

Other vs. Adenocarcinoma:

$$\widehat{\mathrm{OR}}_2 = \frac{\exp[\hat{\alpha}_2 + \hat{\beta}_{21}(1) + \hat{\beta}_{22}(X_2) + \hat{\beta}_{23}(X_3)]}{\exp[\hat{\alpha}_2 + \hat{\beta}_{21}(0) + \hat{\beta}_{22}(X_2) + \hat{\beta}_{23}(X_3)]}$$
$$= \exp\hat{\beta}_{21} = \exp(0.2823) = 1.33$$

Suppose we are interested in the effect of AGEGP, controlling for the effects of ESTROGEN and SMOKING. The odds ratio for the effect of AGEGP in the comparison of Adenosquamous ($D = 1$) to Adenocarcinoma ($D = 0$) is equal to e to the $\hat{\beta}_{11}$ or exp(0.9871) equals 2.68.

The odds ratio for the effect of AGEGP in the comparison of Other type ($D = 2$) to Adenocarcinoma ($D = 0$) is equal to e to the $\hat{\beta}_{21}$ or exp(0.2823) equals 1.33.

Interpretation of ORs

Three-variable vs. one-variable model

Three-variable model:

$\Rightarrow$ AGEGP | ESTROGEN, SMOKING

One-variable model:

$\Rightarrow$ AGEGP | no control variables

Our interpretation of the results for the three-variable model differs from that of the one-variable model. The effect of AGEGP on the outcome is now estimated while controlling for the effects of ESTROGEN and SMOKING.

Odds ratios for effect of AGEGP:

	Model	
Comparison	AGEGP ESTROGEN SMOKING	AGEGP
1 vs. 0	2.68	2.18
2 vs. 0	1.33	1.53

If we compare the model with three predictor variables with the model with only AGEGP included, the effect of AGEGP in the reduced model is weaker for the comparison of Adenosquamous to Adenocarcinoma ($\widehat{\mathrm{OR}} = 2.18$ vs. 2.68), but is stronger for the comparison of Other to Adenocarcinoma ($\widehat{\mathrm{OR}} = 1.53$ vs. 1.33).

Results suggest bias for single-predictor model:

- Toward null for comparison of category 1 vs. 0
- Away from null for comparison of category 2 vs. 0.

These results suggest that estrogen use and smoking status act as confounders of the relationship between age group and the tumor category outcome. The results of the single-predictor model suggest a bias toward the null value (i.e., 1) for the comparison of Adenosquamous to Adenocarcinoma, whereas the results suggest a bias away from the null for the comparison of Other to Adenocarcinoma. These results illustrate that assessment of confounding can have added complexity in the case of multilevel outcomes.

EXAMPLE (*continued*)

95% confidence intervals

Use standard errors from three-variable model:

$$s_{\hat{\beta}_{11}} = 0.4118, \quad s_{\hat{\beta}_{21}} = 0.3280$$

95% CI for OR_1

$= \exp[0.9871 \pm 1.96(0.4118)$

$= (1.20, 6.01)$

95% CI for OR_2

$= \exp[0.2832 \pm 1.96(0.3280)$

$= (0.70, 2.52)$

$\left.\begin{array}{l}\text{Likelihood ratio test}\\ \text{Wald tests}\end{array}\right\}$ same procedures as with one predictor

Likelihood ratio test

	−2 ln L
Reduced:	500.97
Full:	494.41

Difference: 6.56
($\sim \chi^2$, with 2 df)
P-value = 0.04

Wald tests

$H_0: \beta_{11} = 0$ (category 1 vs. 0)

$$Z = \frac{0.9871}{0.4118} = 2.40, \quad P = 0.02$$

$H_0: \beta_{21} = 0$ (category 2 vs. 0)

$$Z = \frac{0.2832}{0.3280} = 0.86, \quad P = 0.39$$

The 95% confidence intervals are calculated using the standard errors of the parameter estimates from the three-variable model, which are 0.4118 and 0.3280 for $\hat{\beta}_{11}$ and $\hat{\beta}_{12}$, respectively.

These confidence intervals are calculated with the usual large-sample formula as shown on the left. For OR_1, this yields a confidence interval of 1.20 to 6.01, whereas for OR_2, this yields a confidence interval of 0.70 to 2.52. The confidence interval for OR_2 contains the null value (i.e., 1.0), whereas the interval for OR_1 does not.

The procedures for the likelihood ratio test and for the Wald tests follow the same format as described earlier for the polytomous model with one independent variable.

The likelihood ratio test compares the reduced model without the age group variable to the full model with the age group variable. This test is distributed approximately chi-square with two degrees of freedom. Minus two times the log likelihood for the reduced model is 500.97, and for the full model, it is 494.41. The difference of 6.56 is statistically significant at the 0.05 level ($P = 0.04$).

The Wald tests are carried out as before, with the same null hypotheses. The Wald statistic for β_{11} is equal to 2.40 and for β_{21} is equal to 0.86. The *P*-value for β_{11} is 0.02, while the *P*-value for β_{21} is 0.39. We therefore reject the null hypothesis for β_{11} but not for β_{21}.

EXAMPLE (*continued*)

Conclusion: Is AGEGP significant?*
- ⇒ Yes: Adenocarcinoma vs. Adenosquamous
- ⇒ No: Other vs. Adenosquamous.

*Controlling for ESTROGEN and SMOKING

Decision: Retain or drop AGEGP from model.

We conclude that AGEGP is statistically significant for the Adenosquamous vs. Adenocarcinoma comparison (category 1 vs. 0), but not for the Other vs. Adenocarcinoma comparison (category 2 vs. 0), controlling for ESTROGEN and SMOKING.

The researcher must make a decision about whether to retain AGEGP in the model. If we are interested in both comparisons, then both betas must be retained, even though only one is statistically significant.

We can also consider interaction terms in a polytomous logistic model.

Adding interaction terms

$D = (0, 1, 2)$

Two independent variables (X_1, X_2)

$$\text{log odds} = \alpha_g + \beta_{g1}X_1 + \beta_{g2}X_2 + \beta_{g3}X_1X_2,$$

where $g = 1, 2$

Consider a disease variable that has three categories ($D = 0, 1, 2$) as in our previous example. Suppose our model includes two independent variables, X_1 and X_2, and that we are interested in the potential interaction between these two variables. The log odds could be modeled as α_1 plus $\beta_{g1}X_1$ plus $\beta_{g2}X_2$ plus $\beta_{g3}X_1X_2$. The subscript g ($g = 1, 2$) indicates which comparison is being made (i.e., category 2 vs. 0, or category 1 vs. 0).

Likelihood ratio test

To test significance of interaction terms

H_0: $\beta_{13} = \beta_{23} = 0$

To test for the significance of the interaction term, a likelihood ratio test with two degrees of freedom can be done. The null hypothesis is that β_{13} equals β_{23} equals zero.

Full model: $\alpha_g + \beta_{g1}X_1 + \beta_{g2}X_2 + \beta_{g3}X_1X_2$

Reduced model: $\alpha_g + \beta_{g1}X_1 + \beta_{g2}X_2$,

where $g = 1, 2$

A full model with the interaction term would be fit and its likelihood compared against a reduced model without the interaction term.

Wald test

To test significance of interaction term at each level

H_0: $\beta_{13} = 0$

H_0: $\beta_{23} = 0$

It is also possible to test the significance of the interaction term at each level with Wald tests. The null hypotheses would be that β_{13} equals zero and that β_{23} equals zero. Recall that both terms must either be retained or dropped.

Extending model to G outcomes

The model also easily extends for outcomes with more than three levels.

Outcome variable has G levels: $(0, 1, 2, \ldots, G-1)$

Assume that the outcome has G levels $(0, 1, 2, \ldots, G-1)$. There are now $G-1$ possible comparisons with the reference category.

$$\ln\left[\frac{\mathrm{P}(D=g\,|\,\mathbf{X})}{\mathrm{P}(D=0\,|\,\mathbf{X})}\right] = \alpha_g + \sum_{i=1}^{k} \beta_{gi} X_i,$$

where $g = 1, 2, \ldots, G-1$

If the reference category is 0, we can define the model in terms of $G-1$ expressions of the following form: the log odds of the probability that the outcome is in category g divided by the probability the outcome is in category 0 equals α_g plus the summation of the k independent variables times their β_g coefficients.

Calculation of ORs and CIs as before

The odds ratios and corresponding confidence intervals for the $G-1$ comparisons of category g to category 0 are calculated in the manner previously described. There are now $G-1$ estimated odds ratios and corresponding confidence intervals, for the effect of each independent variable in the model.

$$\left.\begin{array}{l}\text{Likelihood ratio test}\\ \text{Wald tests}\end{array}\right\}\begin{array}{l}\text{same}\\ \text{procedures}\end{array}$$

The likelihood ratio test and Wald test are also calculated as before.

Likelihood ratio test

$$-2\ln L_{\text{reduced}} - (-2\ln L_{\text{full}}) \sim \chi^2$$

with df = number of parameters set to zero under H_0 ($= G-1$ if $k=1$)

For the likelihood ratio test, we test $G-1$ parameter estimates simultaneously for each independent variable. Thus, for testing one independent variable, we have $G-1$ degrees of freedom for the chi-square test statistic comparing the reduced and full models.

Wald test

$$Z = \frac{\hat{\beta}_{g1}}{s_{\hat{\beta}_{g1}}} \sim N(0,1),$$

where $g = 1, 2, \ldots, G-1$

We can also perform a Wald test to examine the significance of individual betas. We have $G-1$ coefficients that can be tested for each independent variable. As before, the set of coefficients must either be retained or dropped.

VI. Likelihood Function for Polytomous Model

(Section may be omitted.)

We now present the likelihood function for polytomous logistic regression. This section may be omitted without loss of continuity.

Outcome with three levels

We will write the function for an outcome variable with three categories. Once the likelihood is defined for three outcome categories, it can easily be extended to G outcome categories.

Consider probabilities of three outcomes:

$\mathrm{P}(D=0), \mathrm{P}(D=1), \mathrm{P}(D=2)$

We begin by examining the individual probabilities for the three outcomes discussed in our earlier example, that is, the probabilities of the tumor being classified as Adenocarcinoma ($D = 0$), Adenosquamous ($D = 1$), or Other ($D = 2$).

Logistic regression: dichotomous outcome:

$$\mathrm{P}(D=1 \mid \mathbf{X}) = \frac{1}{1+\exp\left[-\left(\alpha + \sum_{i=1}^{k} \beta_i X_i\right)\right]}$$

$$\mathrm{P}(D=0 \mid \mathbf{X}) = 1 - \mathrm{P}(D=1 \mid \mathbf{X})$$

Recall that in logistic regression with a dichotomous outcome variable, we were able to write an expression for the probability that the outcome variable was in category 1, as shown on the left, and for the probability the outcome was in category 0, which is 1 minus the first probability.

Polytomous regression: three-level outcome:

$$\mathrm{P}(D=0 \mid \mathbf{X}) + \mathrm{P}(D=1 \mid \mathbf{X}) + \mathrm{P}(D=2 \mid \mathbf{X}) = 1$$

Similar expressions can be written for a three-level outcome. As noted earlier, the sum of the probabilities for the three outcomes must be equal to 1, the total probability.

$$h_1(\mathbf{X}) = \alpha_1 + \sum_{i=1}^{k} \beta_{1i} X_i$$

$$h_2(\mathbf{X}) = \alpha_2 + \sum_{i=1}^{k} \beta_{2i} X_i$$

To simplify notation, we can let $h_1(\mathbf{X})$ be equal to α_1 plus the summation of the k independent variables times their β_1 coefficients and $h_2(\mathbf{X})$ be equal to α_2 plus the summation of the k independent variables times their β_2 coefficients.

$$\frac{\mathrm{P}(D=1 \mid \mathbf{X})}{\mathrm{P}(D=0 \mid \mathbf{X})} = \exp[h_1(\mathbf{X})]$$

$$\frac{\mathrm{P}(D=2 \mid \mathbf{X})}{\mathrm{P}(D=0 \mid \mathbf{X})} = \exp[h_2(\mathbf{X})]$$

The probability for the outcome being in category 1 divided by the probability for the outcome being in category 0 is modeled as e to the $h_1(\mathbf{X})$ and the ratio of probabilities for category 2 and category 0 is modeled as e to the $h_2(\mathbf{X})$.

Solve for $\mathrm{P}(D = 1 \mid \mathbf{X})$ and $\mathrm{P}(D = 2 \mid \mathbf{X})$ in terms of $\mathrm{P}(D = 0 \mid \mathbf{X})$.

Rearranging these equations allows us to solve for the probability that the outcome is in category 1, and for the probability that the outcome is in category 2, in terms of the probability that the outcome is in category 0.

$$\mathrm{P}(D = 1 \mid \mathbf{X}) = \mathrm{P}(D = 0 \mid \mathbf{X}) \exp[h_1(\mathbf{X})]$$
$$\mathrm{P}(D = 2 \mid \mathbf{X}) = \mathrm{P}(D = 0 \mid \mathbf{X}) \exp[h_2(\mathbf{X})]$$

The probability that the outcome is in category 1 is equal to the probability that the outcome is in category 0 times e to the $h_1(\mathbf{X})$. Similarly, the probability that the outcome is in category 2 is equal to the probability that the outcome is in category 0 times e to the $h_2(\mathbf{X})$.

$$\mathrm{P}(D = 0 \mid \mathbf{X}) + \mathrm{P}(D = 0 \mid \mathbf{X}) \exp[h_1(\mathbf{X})] + \mathrm{P}(D = 0 \mid \mathbf{X}) \exp[h_2(\mathbf{X})] = 1$$

These quantities can be substituted into the total probability equation and summed to 1.

Factoring out $\mathrm{P(D} = 0|\mathbf{X})$:

$$\mathrm{P}(D = 0 \mid \mathbf{X})[1 + \exp h_1(\mathbf{X}) + \exp h_2(\mathbf{X})] = 1$$

With some algebra, we find that

$$\mathrm{P}(D = 0 \mid \mathbf{X}) = \frac{1}{1 + \exp[h_1(\mathbf{X})] + \exp[h_2(\mathbf{X})]}$$

With some simple algebra, we can see that the probability that the outcome is in category 0 is 1 divided by the quantity 1 plus e to the $h_1(\mathbf{X})$ plus e to the $h_2(\mathbf{X})$.

and that

$$\mathrm{P}(D = 1 \mid \mathbf{X}) = \frac{\exp[h_1(\mathbf{X})]}{1 + \exp[h_1(\mathbf{X})] + \exp[h_2(\mathbf{X})]}$$

Substituting this value into our earlier equation for the probability that the outcome is in category 1, we obtain the probability that the outcome is in category 1 as e to the $h_1(\mathbf{X})$ divided by one plus e to the $h_1(\mathbf{X})$ plus e to the $h_2(\mathbf{X})$.

and that

$$\mathrm{P}(D = 2 \mid \mathbf{X}) = \frac{\exp[h_2(\mathbf{X})]}{1 + \exp[h_1(\mathbf{X})] + \exp[h_2(\mathbf{X})]}$$

The probability that the outcome is in category 2 can be found in a similar way, as shown on the left.

$L \Leftrightarrow$ joint probability of observed data.

The ML method chooses parameter estimates that maximize L

Recall that the likelihood function (L) represents the joint probability of observing the data that have been collected and that the method of maximum likelihood (ML) chooses that estimator of the set of unknown parameters that maximizes the likelihood.

Subjects: $j = 1, 2, 3, \ldots, n$

$$y_{j0} = \begin{cases} 1 & \text{if outcome} = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$y_{j1} = \begin{cases} 1 & \text{if outcome} = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$y_{j2} = \begin{cases} 1 & \text{if outcome} = 2 \\ 0 & \text{otherwise} \end{cases}$$

Assume that there are n subjects in the dataset, numbered from $j = 1$ to n. If the outcome for subject j is in category 0, then we let an indicator variable, y_{j0}, be equal to 1, otherwise y_{j0} is equal to 0. We similarly create indicator variables y_{j1} and y_{j2} to indicate whether the subject's outcome is in category 1 or category 2.

$$\text{P}(D = 0 \,|\, \mathbf{X})^{y_{j0}} \ \text{P}(D = 1 \,|\, \mathbf{X})^{y_{j1}} \\ \times \text{P}(D = 2 \,|\, \mathbf{X})^{y_{j2}}$$

The contribution of each subject to the likelihood is the probability that the outcome is in category 0, raised to the y_{j0} power, times the probability that the outcome is in category 1, raised to the y_{j1}, times the probability that the outcome is in category 2, raised to the y_{j2}.

$$y_{j0} + y_{j1} + y_{j2} = 1$$

since each subject has one outcome

Note that each individual subject contributes to only one of the category probabilities, since only one of the indicator variables will be nonzero.

$$\prod_{j=1}^{n} \text{P}(D=0 \,|\, \mathbf{X})^{y_{j0}} \text{P}(D=1 \,|\, \mathbf{X})^{y_{j1}} \text{P}(D=2 \,|\, \mathbf{X})^{y_{j2}}$$

The joint probability for the likelihood is the product of all the individual subject probabilities, assuming subject outcomes are independent.

Likelihood for G outcome categories:

$$\prod_{j=1}^{n} \prod_{g=0}^{G-1} \text{P}(D = g \,|\, \mathbf{X})^{y_{jg}},$$

where

$$y_{jg} = \begin{cases} 1 \text{ if the } j\text{th subject has } D = g \\ \quad (g = 0, 1, \ldots, G-1) \\ 0 \text{ if otherwise} \end{cases}$$

The likelihood can be generalized to include G outcome categories by taking the product of each individual's contribution across the G outcome categories.

Estimated αs and βs are those which maximize L

The unknown parameters that will be estimated by maximizing the likelihood are the alphas and betas in the probability that the disease outcome is in category g, where g equals $0, 1, \ldots, G - 1$.

VII. Polytomous vs. Multiple Standard Logistic Regressions

Polytomous vs. separate logistic models

One may wonder how using a polytomous model compares with using two or more separate dichotomous logistic models.

Polytomous model uses data on all outcome categories in *L*.

Separate standard logistic model uses data ononly two outcome categories at a time.

⇓

Parameter and variance estimates may differ.

The likelihood function for the polytomous model utilizes the data involving all categories of the outcome variable in a single structure. In contrast, the likelihood function for a dichotomous logistic model utilizes the data involving only two categories of the outcome variable. In other words, different likelihood functions are used when fitting each dichotomous model separately than when fitting a polytomous model that considers all levels simultaneously. Consequently, both the estimation of the parameters and the estimation of the variances of the parameter estimates may differ when comparing the results from fitting separate dichotomous models to the results from the polytomous model.

Special case: One dichotomous predictor Polytomous and standard logistic models ⇒ same estimates

In the special case of a polytomous model with one dichotomous predictor, fitting separate logistic models yields the same parameter estimates and variance estimates as fitting the polytomous model.

VIII. SUMMARY

✓ Chapter 9: Polytomous Logistic Regression

This presentation is now complete. We have described a method of analysis, polytomous regression, for the situation where the outcome variable has more than two categories.

We suggest that you review the material covered here by reading the detailed outline that follows. Then, do the practice exercises and test.

Chapter 10: Ordinal Logistic Regression

If there is no inherent ordering of the outcome categories, a polytomous regression model is appropriate. If there is an inherent ordering of the outcome categories, then an ordinal logistic regression model may also be appropriate. The proportional odds model is one such ordinal model, which may be used if the proportional odds assumption is met. This model is discussed in Chap. 10.

Detailed Outline

I. Overview (pages 432–433)

A. Focus: modeling outcomes with more than two levels.

B. Using previously described techniques by combining outcome categories.

C. Nominal vs. ordinal outcomes.

II. Polytomous logistic regression: An example with three categories (pages 434–437)

A. Nominal outcome: variable has no inherent order.

B. Consider "odds-like" expressions, which are ratios of probabilities.

C. Example with three categories and one predictor (X_1):

$$\ln\left[\frac{P(D=1\,|\,X_1)}{P(D=0\,|\,X_1)}\right]=\alpha_1+\beta_{11}X_1,$$

$$\ln\left[\frac{P(D=2\,|\,X_1)}{P(D=0\,|\,X_1)}\right]=\alpha_2+\beta_{21}X_1.$$

III. Odds ratio with three categories (pages 437–441)

A. Computation of OR in polytomous regression is analogous to standard logistic regression, except that there is a separate odds ratio for each comparison.

B. The general formula for the odds ratio for any two levels of the exposure variable (X_1^{**} and X_1^{*}) in a no-interaction model is

$$\text{OR}_g=\exp\left[(\beta_{g1}(X_1^{**}-X_1^{*})\right],\quad \text{where } g=1,2.$$

IV. Statistical inference with three categories (pages 441–444)

A. Two types of statistical inferences are often of interest in polytomous regression:

i. Testing hypotheses

ii. Deriving interval estimates

B. Confidence interval estimation is analogous to standard logistic regression.

C. The general large-sample formula (no-interaction model) for a 95% confidence interval for comparison of outcome level g vs. the reference category, for any two levels of the independent variable (X_1^{**} and X_1^{*}), is

$$\exp\left\{\hat\beta_{g1}(X_1^{**}-X_1^{*})\pm 1.96(X_1^{**}-X_1^{*})s_{\hat\beta_{g1}}\right\}.$$

D. The likelihood ratio test is used to test hypotheses about the significance of the predictor variable(s).

i. With three levels of the outcome variable, there are two comparisons and two estimated coefficients for each predictor

ii. The null hypothesis is that each of the 2 beta coefficients (for a given predictor) is equal to zero

iii. The test compares the log likelihood of the full model with the predictor to that of the reduced model without the predictor. The test is distributed approximately chi-square, with 2 df for each predictor tested

E. The Wald test is used to test the significance of the predictor at a single outcome level. The procedure is analogous to standard logistic regression.

V. Extending the polytomous model to *G* outcomes and *k* predictors (pages 444–449)

A. The model easily extends to include k independent variables.

B. The general form of the model for G outcome levels is

$$\ln\left[\frac{\mathrm{P}(D=g\,|\,\mathrm{X})}{\mathrm{P}(D=0\,|\,\mathrm{X})}\right]=\alpha_g+\sum_{i=1}^{k}\beta_{gi}X_i,$$

where $g=1,2,\ldots,G-1$.

C. The calculation of the odds ratio, confidence intervals, and hypothesis testing using the likelihood ratio and Wald tests remains the same.

D. Interaction terms can be added and tested in a manner analogous to standard logistic regression.

VI. Likelihood function for polytomous model (pages 450–452)

A. For an outcome variable with G categories, the likelihood function is

$$\prod_{j=1}^{n}\prod_{g=0}^{G-1}\mathrm{P}(D=g\,|\,\mathbf{X})^{y_{jg}},\quad\text{where}$$

$$y_{jg}=\begin{cases}1 & \text{if the } j\text{th subject has } D=g\\ 0 & \text{if otherwise}\end{cases}$$

where n is the total number of subjects and $g=0,1,\ldots,G-1$.

VII. Polytomous vs. multiple standard logistic regressions (page 453)

A. The likelihood for polytomous regression takes into account all of the outcome categories; the likelihood for the standard logistic model considers only two outcome categories at a time.

B. Parameter and standard error estimates may differ.

VIII. Summary (page 453)

Practice Exercises

Suppose we are interested in assessing the association between tuberculosis and degree of viral suppression in HIV-infected individuals on antiretroviral therapy, who have been followed for 3 years in a hypothetical cohort study. The outcome, tuberculosis, is coded as none ($D = 0$), latent ($D = 1$), or active ($D = 2$). The degree of viral suppression (VIRUS) is coded as undetectable (VIRUS = 0) or detectable (VIRUS = 1). Previous literature has shown that it is important to consider whether the individual has progressed to AIDS (no = 0, yes = 1), and is compliant with therapy (COMPLIANCE: no = 1, yes = 0). In addition, AGE (continuous) and GENDER (female = 0, male = 1) are potential confounders. Also, there may be interaction between progression to AIDS and compliance with therapy (AIDSCOMP = AIDS × COMPLIANCE).

We decide to run a polytomous logistic regression to analyze these data. Output from the regression is shown below. (The results are hypothetical.) The reference category for the polytomous logistic regression is no tuberculosis ($D = 0$). This means that a descending option was used to obtain the polytomous regression output for the model, so Intercept 1 (and the coefficient estimates that follow) pertains to the comparison of $D = 2$ to $D = 0$, and Intercept 2 pertains to the comparison of $D = 1$ to $D = 0$.

Variable	Coefficient	S.E.
Intercept 1	−2.82	0.23
VIRUS	1.35	0.11
AIDS	0.94	0.13
COMPLIANCE	0.49	0.21
AGE	0.05	0.04
GENDER	0.41	0.22
AIDSCOMP	0.33	0.14
Intercept 2	−2.03	0.21
VIRUS	0.95	0.14
AIDS	0.76	0.15
COMPLIANCE	0.34	0.17
AGE	0.03	0.03
GENDER	0.25	0.18
AIDSCOMP	0.31	0.17

1. State the form of the polytomous model in terms of variables and unknown parameters.
2. For the above model, state the fitted model in terms of variables and estimated coefficients.
3. Is there an assumption with this model that the outcome categories are ordered? Is such an assumption reasonable?

4. Compute the estimated odds ratio for a 25-year-old noncompliant male, with a detectable viral load, who has progressed to AIDS, compared with a similar female. Consider the outcome comparison latent tuberculosis vs. none ($D = 1$ vs. $D = 0$).
5. Compute the estimated odds ratio for a 25-year-old noncompliant male, with a detectable viral load, who has progressed to AIDS, compared with a similar female. Consider the outcome comparison active tuberculosis vs. none ($D = 2$ vs. $D = 0$).
6. Use the results from the previous two questions to obtain an estimated odds ratio for a 25-year-old noncompliant male, with a detectable viral load, who has progressed to AIDS, compared with a similar female, with the outcome comparison active tuberculosis vs. latent tuberculosis ($D = 2$ vs. $D = 1$).

Note. If the same polytomous model was run with latent tuberculosis designated as the reference category ($D = 1$), the output could be used to directly estimate the odds ratio comparing a male to a female with the outcome comparison active tuberculosis vs. latent tuberculosis ($D = 2$ vs. $D = 1$). This odds ratio can also indirectly be estimated with $D = 0$ as the reference category. This is justified since the OR ($D = 2$ vs. $D = 0$) divided by the OR ($D = 1$ vs. $D = 0$) equals the OR ($D = 2$ vs. $D = 1$). However, if each of these three odds ratios were estimated with three separate logistic regressions, then the three estimated odds ratios are not generally so constrained since the three outcomes are not modeled simultaneously.

7. Use Wald statistics to assess the statistical significance of the interaction of AIDS and COMPLIANCE in the model at the 0.05 significance level.
8. Estimate the odds ratio(s) comparing a subject who has progressed to AIDS to one who has not, with the outcome comparison active tuberculosis vs. none ($D = 2$ vs. $D = 0$), controlling for viral suppression, age, and gender.
9. Estimate the odds ratio with a 95% confidence interval for the viral load suppression variable (detectable vs. undetectable), comparing active tuberculosis to none, controlling for the effect of the other covariates in the model.
10. Estimate the odds of having latent tuberculosis vs. none ($D = 1$ vs. $D = 0$) for a 20-year-old compliant female, with an undetectable viral load, who has not progressed to AIDS.

Test

True or False (Circle T or F)

T F 1. An outcome variable with categories North, South, East, and West is an ordinal variable.

T F 2. If an outcome has three levels (coded 0, 1, 2), then the ratio of P(D = 1)/P(D = 0) can be considered an odds if the outcome is conditioned on only the two outcome categories being considered (i.e., $D = 1$ and $D = 0$).

T F 3. In a polytomous logistic regression in which the outcome variable has five levels, there will be four intercepts.

T F 4. In a polytomous logistic regression in which the outcome variable has five levels, each independent variable will have one estimated coefficient.

T F 5. In a polytomous model, the decision of which outcome category is designated as the reference has no bearing on the parameter estimates since the choice of reference category is arbitrary.

6. Suppose the following polytomous model is specified for assessing the effects of AGE (coded continuously), GENDER (male = 1, female = 0), SMOKE (smoker = 1, nonsmoker = 0), and hypertension status (HPT) (yes = 1, no = 0) on a disease variable with four outcomes (coded $D = 0$ for none, $D = 1$ for mild, $D = 2$ for severe, and $D = 3$ for critical).

$$\ln\left[\frac{\mathrm{P}(D=g\,|\,\mathbf{X})}{\mathrm{P}(D=0\,|\,\mathbf{X})}\right] = \alpha_g + \beta_{g1}\,\mathrm{AGE} + \beta_{g2}\,\mathrm{GENDER} + \beta_{g3}\,\mathrm{SMOKE} + \beta_{g4}\,\mathrm{HPT},$$

where $g = 1, 2, 3$.

Use the model to give an expression for the odds (severe vs. none) for a 40-year-old nonsmoking male. (*Note.* Assume that the expression [P($D = g\,|\,\mathbf{X}$ / P($D = 0\,|\,\mathbf{X}$)] gives the odds for comparing group g with group 0, even though this ratio is not, strictly speaking, an odds.)

7. Use the model in Question 6 to obtain the odds ratio for male vs. female, comparing mild disease to none, while controlling for AGE, SMOKE, and HPT.

8. Use the model in Question 6 to obtain the odds ratio for a 50-year-old vs. a 20-year-old subject, comparing severe disease to none, while controlling for GENDER, SMOKE, and HPT.

9. For the model in Question 6, describe how you would perform a likelihood ratio test to simultaneously test the significance of the SMOKE and HPT coefficients.

State the null hypothesis, the test statistic, and the distribution of the test statistic under the null hypothesis.

10. Extend the model from Question 6 to allow for interaction between AGE and GENDER and between SMOKE and GENDER. How many additional parameters would be added to the model?

Answers to Practice Exercises

1. Polytomous model:

$$\ln\left[\frac{\mathrm{P}(D=g\mid\mathbf{X})}{\mathrm{P}(D=0\mid\mathbf{X})}\right]=\alpha_g+\beta_{g1}\mathrm{VIRUS}+\beta_{g2}\mathrm{AIDS}+\beta_{g3}\mathrm{COMPLIANCE}+\beta_{g4}\mathrm{AGE}$$
$$+\beta_{g5}\mathrm{GENDER}+\beta_{g6}\mathrm{AIDSCOMP},$$

where $g = 1, 2$.

2. Polytomous fitted model:

$$\ln\left[\frac{\hat{\mathrm{P}}(D=2\mid\mathbf{X})}{\hat{\mathrm{P}}(D=0\mid\mathbf{X})}\right]=-2.82+1.35\mathrm{VIRUS}+0.94\mathrm{AIDS}+0.49\mathrm{COMPLIANCE}$$
$$+0.05\mathrm{AGE}+0.41\mathrm{GENDER}+0.33\mathrm{AIDSCOMP},$$

$$\ln\left[\frac{\hat{\mathrm{P}}(D=1\mid\mathbf{X})}{\hat{\mathrm{P}}(D=0\mid\mathbf{X})}\right]=-2.03+0.95\mathrm{VIRUS}+0.76\mathrm{AIDS}+0.34\mathrm{COMPLIANCE}$$
$$+0.03\mathrm{AGE}+0.25\mathrm{GENDER}+0.31\mathrm{AIDSCOMP}.$$

3. No, the polytomous model does not assume an ordered outcome. The categories given do have a natural order however, so that an ordinal model may also be appropriate (see Chap. 10).
4. $\widehat{\mathrm{OR}}_{1\mathrm{vs}0}=\exp(0.25)=1.28.$
5. $\widehat{\mathrm{OR}}_{2\mathrm{vs}0}=\exp(0.41)=1.51.$
6. $\widehat{\mathrm{OR}}_{2\mathrm{vs}1}=\exp(0.41)/\exp(0.25)=\exp(0.16)=1.17.$
7. Two Wald statistics:

$$H_0\colon \beta_{16}=0;\quad z_1=\frac{0.31}{0.17}=1.82;\ \text{two-tailed } P\text{-value}: 0.07,$$
$$H_0\colon \beta_{26}=0;\quad z_2=\frac{0.33}{0.14}=2.36;\ \text{two-tailed } P\text{-value}: 0.02.$$

The P-value is statistically significant at the 0.05 level for the hypothesis $\beta_{26}=0$ but not for the hypothesis $\beta_{16}=0$. Since we must either keep or drop both interaction parameters from the model, we elect to keep both parameters because there is a suggestion of interaction between AIDS and COMPLIANCE. Alternatively, a likelihood ratio test could be performed. The likelihood ratio test has the advantage that only one test statistic needs to be calculated.

8. Estimated odds ratios (AIDS progression: yes vs. no):

$$\text{for COMPLIANCE} = 0\colon \exp(0.94) = 2.56,$$
$$\text{for COMPLIANCE} = 1\colon \exp(0.94 + 0.33) = 3.56.$$

9. $\widehat{\text{OR}} = \exp(1.35) = 3.86; 95\% \text{ CI}\colon \exp[1.35 \pm 1.96(0.11)]$
   $= (3.11, 4.79).$

10. Estimated odds $= \exp[-2.03 + (0.03)(20)]$
    $= \exp(-1.43) = 0.24.$

13 Ordinal Logistic Regression

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_13, © Springer Science+Business Media, LLC 2010

Introduction

In this chapter, the standard logistic model is extended to handle outcome variables that have more than two ordered categories. When the categories of the outcome variable have a natural order, ordinal logistic regression may be appropriate.

The mathematical form of one type of ordinal logistic regression model, the proportional odds model, and its interpretation are developed. The formulas for the odds ratio and confidence intervals are derived, and techniques for testing hypotheses and assessing the statistical significance of independent variables are shown.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

I. Overview (page 466)

II. Ordinal logistic regression: The proportional odds model (pages 466–472)

III. Odds ratios and confidence limits (pages 472–475)

IV. Extending the ordinal model (pages 476–478)

V. Likelihood function for ordinal model (pages 478–479)

VI. Ordinal vs. multiple standard logistic regressions (pages 479–481)

VII. Summary (page 481)

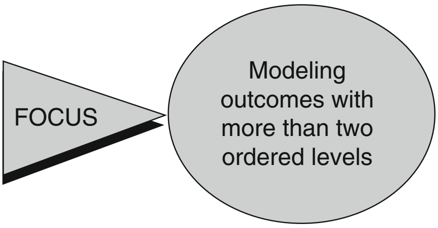
Objectives

Upon completing this chapter, the learner should be able to:

1. State or recognize when the use of ordinal logistic regression may be appropriate.
2. State or recognize the proportional odds assumption.
3. State or recognize the proportional odds model.
4. Given a printout of the results of a proportional odds model:
   a. State the formula and compute the odds ratio.
   b. State the formula and compute a confidence interval for the odds ratio.
   c. Test hypotheses about the model parameters using the likelihood ratio test or the Wald test, stating the null hypothesis and the distribution of the test statistic with the corresponding degrees of freedom under the null hypothesis.

Presentation

I. Overview



This presentation and the presentation in Chap. 12 describe approaches for extending the standard logistic regression model to accommodate a disease, or outcome, variable that has more than two categories. The focus of this presentation is on modeling outcomes with more than two *ordered* categories. We describe the *form* and key *characteristics* of one model for such outcome variables: ordinal logistic regression using the proportional odds model.

Ordinal: levels have natural ordering

EXAMPLE

Tumor grade:

- Well differentiated
- Moderately differentiated
- Poorly differentiated

Ordinal variables have a natural ordering among the levels. An example is cancer tumor grade, ranging from well differentiated to moderately differentiated to poorly differentiated tumors.

Ordinal outcome ⇒ Polytomous model or ordinal model

An ordinal outcome variable with three or more categories can be modeled with a polytomous model, as discussed in Chap. 12, but can also be modeled using ordinal logistic regression, provided that certain assumptions are met.

Ordinal model takes into account order of outcome levels

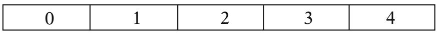
Ordinal logistic regression, unlike polytomous regression, takes into account any inherent ordering of the levels in the disease or outcome variable, thus making fuller use of the ordinal information.

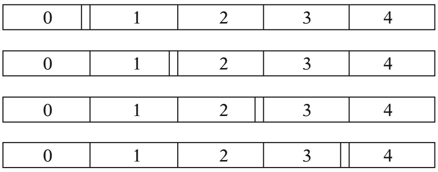
II. Ordinal Logistic Regression: The Proportional Odds Model

Proportional Odds Model/ Cumulative Logit Model

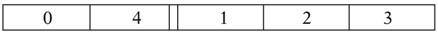
The ordinal logistic model that we shall develop is called the proportional odds or cumulative logit model.

Illustration





But, cannot allow



To illustrate the proportional odds model, assume we have an outcome variable with five categories and consider the four possible ways to divide the five categories into two collapsed categories preserving the natural order.

We could compare category 0 to categories 1 through 4, or categories 0 and 1 to categories 2 through 4, or categories 0 through 2 to categories 3 and 4, or, finally, categories 0 through 3 to category 4. However, we could not combine categories 0 and 4 for comparison with categories 1, 2, and 3, since that would disrupt the natural ordering from 0 through 4.

For G categories ⇒ $G-1$ ways to dichotomize outcome:

$D \geq 1$ vs. $D < 1$;
$D \geq 2$ vs. $D < 2, \ldots,$
$D \geq G-1$ vs. $D < G-1$

$$\text{odds } (D \geq g) = \frac{P(D \geq g)}{P(D < g)},$$

where $g = 1, 2, 3, \ldots, G-1$

More generally, if an ordinal outcome variable D has G categories ($D = 0, 1, 2, \ldots, G-1$), then there are $G-1$ ways to dichotomize the outcome: ($D \geq 1$ vs. $D < 1$; $D \geq 2$ vs. $D < 2, \ldots,$ $D \geq G-1$ vs. $D < G-1$). With this categorization of D, the odds that $D \geq g$ is equal to the probability of $D \geq g$ divided by the probability of $D < g$, where ($g = 1, 2, 3, \ldots, G-1$).

Proportional odds assumption

EXAMPLE

OR ($D \geq 1$) = OR ($D \geq 4$)
Comparing two exposure groups
e.g., $E = 1$ vs. $E = 0$,

where

$$\text{OR}_{(D \geq 1)} = \frac{\text{odds}[(D \geq 1) \mid E = 1]}{\text{odds}[(D \geq 1) \mid E = 0]}$$

$$\text{OR}_{(D \geq 4)} = \frac{\text{odds}[(D \geq 4) \mid E = 1]}{\text{odds}[(D \geq 4) \mid E = 0]}$$

The proportional odds model makes an important assumption. Under this model, the odds ratio assessing the effect of an exposure variable for any of these comparisons will be the same regardless of where the cut-point is made. Suppose we have an outcome with five levels and one dichotomous exposure ($E = 1$, $E = 0$). Then, under the proportional odds assumption, the odds ratio that compares categories greater than or equal to 1 to less than 1 is the same as the odds ratio that compares categories greater than or equal to 4 to less than 4.

Same odds ratio regardless of where categories are dichotomized

In other words, the odds ratio is *invariant* to where the outcome categories are dichotomized.

Ordinal

Variable	Parameter
Intercept	$\alpha_1, \alpha_2, \ldots, \alpha_{G-1}$
X_1	β_1

Polytomous

Variable	Parameter
Intercept	$\alpha_1, \alpha_2, \ldots, \alpha_{G-1}$
X_1	$\beta_{11}, \beta_{21}, \ldots, \beta_{(G-1)1}$

This implies that if there are G outcome categories, there is only one parameter (β) for each of the predictors variables (e.g., β_1 for predictor X_1). However, there is still a separate intercept term (α_g) for each of the $G-1$ comparisons.

This contrasts with polytomous logistic regression, where there are $G-1$ parameters for each predictor variable, as well as a separate intercept for each of the $G-1$ comparisons.

Odds are *not* invariant

EXAMPLE

$\text{odds}(D \geq 1) \neq \text{odds}(D \geq 4)$

where, for $E = 0$,

$$\text{odds}(D \geq 1) = \frac{P(D \geq 1 \mid E = 0)}{P(D < 1 \mid E = 0)}$$

$$\text{odds}(D \geq 4) = \frac{P(D \geq 4 \mid E = 0)}{P(D < 4 \mid E = 0)}$$

but

$\text{OR}(D \geq 1) = \text{OR}(D \geq 4)$

The assumption of the invariance of the odds ratio regardless of cut-point is *not* the same as assuming that the *odds* for a given exposure pattern is invariant. Using our previous example, for a given exposure level E (e.g., $E = 0$), the odds comparing categories greater than or equal to 1 to less than 1 does *not* equal the odds comparing categories greater than or equal to 4 to less than 4.

Proportional odds model: *G* outcome levels and one predictor (*X*)

$$P(D \geq g \mid X_1) = \frac{1}{1 + \exp[-(\alpha_g + \beta_1 X_1)]},$$

where $g = 1, 2, \ldots, G-1$

We now present the form for the proportional odds model with an outcome (D) with G levels ($D = 0, 1, 2, \ldots, G-1$) and one independent variable (X_1). The probability that the disease outcome is in a category greater than or equal to g, given the exposure, is 1 over 1 plus e to the negative of the quantity α_g plus $\beta_1 X_1$.

$$\begin{aligned} 1 - P(D \geq g \mid \mathbf{X}_1) &= 1 - \frac{1}{1 + \exp[-(\alpha_g + \beta_1 \mathrm{X}_1)]} \\ &= \frac{\exp[-(\alpha_g + \beta_1 X_1)]}{1 + \exp[-(\alpha_g + \beta_1 X_1)]} \\ &= P(D < g \mid X_1) \end{aligned}$$

The probability that the disease outcome is in a category *less* than g is equal to 1 minus the probability that the disease outcome is greater than or equal to category g.

Equivalent model definition

$$\text{odds} = \frac{\mathrm{P}(D \geq g \,|\, X_1)}{1 - \mathrm{P}(D \geq g \,|\, X_1)} = \frac{\mathrm{P}(D \geq g \,|\, X_1)}{\mathrm{P}(D < g \,|\, X_1)}$$

$$= \frac{\dfrac{1}{1 + \exp[-(\alpha_g + \beta_1 X_1)]}}{\dfrac{\exp[-(\alpha_g + \beta_1 X_1)]}{1 + \exp[-(\alpha_g + \beta_1 X_1)]}}$$

$$= \exp(\alpha_g + \beta_1 X_1)$$

The model can be defined equivalently in terms of the odds of an inequality. If we substitute the formula $\mathrm{P}(D \geq g \,|\, X_1)$ into the expression for the odds and then perform some algebra (as shown on the left), we find that the *odds* is equal to e to the quantity α_g plus $\beta_1 X_1$.

Proportional odds model:	vs.	Standard logistic model:
$\mathrm{P}(D \geq g \,|\, \mathbf{X})$		$\mathrm{P}(D = g \,|\, \mathbf{X})$

The proportional odds model is written differently from the standard logistic model. The model is formulated as the probability of an inequality, that is, that the outcome D is greater than or equal to g.

Proportional odds model:	vs.	Polytomous model:
β_1		β_{g1}
↗		↗
no g subscript		g subscript

The model also differs from the polytomous model in an important way. The beta is not subscripted by g. This is consistent with the proportional odds assumption that only one parameter is required for each independent variable.

Alternate model formulation:

key differences

$$\text{odds} = \frac{\mathrm{P}(D^* \leq g \,|\, X_1)}{\mathrm{P}(D^* > g \,|\, X_1)} = \exp(\alpha_g^* - \beta_1^* X_1),$$

where $g = 1, 2, 3, \ldots, G{-}1$
and $D^* = 1, 2, \ldots, G$

Comparing formulations

$$\beta_1 = \beta_1^*$$
$$\text{but } \alpha_g = -\alpha_g^*$$

An alternate formulation of the proportional odds model is to define the model as the odds of D^* less than or equal to g given the exposure is equal to e to the quantity $\alpha_g^* - \beta_1^* X_1$, where $g = 1, 2, 3, \ldots, G{-}1$ and where $D^* = 1, 2, \ldots, G$. The two key differences with this formulation are the direction of the inequality ($D^* \leq g$) and the negative sign before the parameter β_1^*. In terms of the beta coefficients, these two key differences "cancel out" so that $\beta_1 = \beta_1^*$. Consequently, if the same data are fit for each formulation of the model, the same parameter estimates of beta would be obtained for each model. However, the intercepts for the two formulations differ as $\alpha_g = -\alpha_g^*$.

Formulation affects computer output

- SAS: consistent with first
- SPSS and Stata: consistent with alternative formulation

We have presented two ways of parameterizing the model because different software packages can present slightly different output depending on the way the model is formulated. SAS software presents output consistent with the way we have formulated the model, whereas SPSS and Stata software present output consistent with the alternate formulation (see Appendix).

Advantage of ($D \geq g$):

Consistent with formulations of standard logistic and polytomous models

$\Downarrow$

For 2-level outcome ($D = 0, 1$), all three reduce to same model.

An advantage to our formulation of the model (i.e., in terms of the odds of $D \geq g$) is that it is consistent with the way that the standard logistic model and polytomous logistic model are presented. In fact, for a two-level outcome (i.e., $D = 0, 1$), the standard logistic, polytomous, and ordinal models reduce to the same model. However, the alternative formulation is consistent with the way the model has historically often been presented (McCullagh, 1980). Many models can be parameterized in different ways. This need not be problematic as long as the investigator understands how the model is formulated and how to interpret its parameters.

EXAMPLE

Black/White Cancer Survival Study

$$\mathbf{E} = \text{RACE}\begin{cases} 0 & \text{if white} \\ 1 & \text{if black} \end{cases}$$

$$\mathbf{D} = \text{GRADE}\begin{cases} 0 & \text{if well differentiated} \\ 1 & \text{if moderately differentiated} \\ 2 & \text{if poorly differentiated} \end{cases}$$

Next, we present an example of the proportional odds model using data from the Black/White Cancer Survival Study (Hill et al., 1995). Suppose we are interested in assessing the effect of RACE on tumor grade among women with invasive endometrial cancer. RACE, the exposure variable, is coded 0 for white and 1 for black. The disease variable, tumor grade, is coded 0 for well-differentiated tumors, 1 for moderately differentiated tumors, and 2 for poorly differentiated tumors.

Ordinal: Coding of disease meaningful
Polytomous: Coding of disease arbitrary

Here, the coding of the disease variable reflects the ordinal nature of the outcome. For example, it is necessary that moderately differentiated tumors be coded between poorly differentiated and well-differentiated tumors. This contrasts with polytomous logistic regression, in which the order of the coding is not reflective of an underlying order in the outcome variable.

EXAMPLE (*continued*)

	White (0)	Black (1)
Well differentiated	104	26
Moderately differentiated	72	33
Poorly differentiated	31	22

The 3 × 2 table of the data is presented on the left.

A simple check of the proportional odds assumption:

In order to examine the proportional odds assumption, the table is collapsed to form two other tables.

	White	Black
Well + moderately differentiated	176	59
Poorly differentiated	31	22

$\widehat{\text{OR}} = 2.12$

The first table combines the well-differentiated and moderately differentiated levels. The odds ratio is 2.12.

	White	Black
Well differentiated	104	26
Moderately + poorly differentiated	103	55

$\widehat{\text{OR}} = 2.14$

The second table combines the moderately and poorly differentiated levels. The odds ratio for this data is 2.14.

The odds ratios from the two collapsed tables are similar and thus provide evidence that the proportional odds assumption is not violated. It would be unusual for the collapsed odds ratios to match perfectly. The odds ratios do not have to be exactly equal; as long as they are "close", the proportional odds assumption may be considered reasonable.

Requirement: Collapsed ORs should be "close"

	$E = 0$	$E = 1$
$D = 0$	45	30
$D = 1$	40	15
$D = 2$	50	60

Here is a different 3 × 2 table. This table will be collapsed in a similar fashion as the previous one.

	$E = 0$	$E = 1$
$D = 0 + 1$	85	45
$D = 2$	50	60

$\widehat{\text{OR}} = 2.27$

	$E = 0$	$E=1$
$D = 0$	45	30
$D = 1+2$	90	75

$\widehat{\text{OR}} = 1.25$

The two collapsed tables are presented on the left. The odds ratios are 2.27 and 1.25. In this case, we would question whether the proportional odds assumption is appropriate, since one odds ratio is nearly twice the value of the other.

Statistical test of assumption: *Score test*

Compares ordinal vs. polytomous models

Test statistic $\sim \chi^2$ under H_0 with df = number of OR parameters tested

There is also a statistical test – a **Score test** – designed to evaluate whether a model constrained by the proportional odds assumption (i.e., an ordinal model) is significantly different from the corresponding model in which the odds ratio parameters are not constrained by the proportional odds assumption (i.e., a polytomous model). The test statistic is distributed approximately chi-square, with degrees of freedom equal to the number of odds ratio parameters being tested.

Alternate models for ordinal data:

- Continuation ratio
- Partial proportional odds
- Stereotype regression

If the proportional odds assumption is inappropriate, there are other ordinal logistic models that may be used that make alternative assumptions about the ordinal nature of the outcome. Examples include a continuation ratio model, a partial proportional odds model, and stereotype regression models. These models are beyond the scope of the current presentation. [See the review by Ananth and Kleinbaum (1997)].

III. Odds Ratios and Confidence Limits

ORs: same method as SLR to compute ORs.

After the proportional odds model is fit and the parameters estimated, the process for computing the odds ratio is the same as in standard logistic regression (SLR).

Special case: one independent variable

$X_1 = 1$ or $X_1 = 0$

$$\begin{aligned}\text{odds}(D \geq g) &= \frac{\text{P}(D \geq g\,|\,X_1)}{\text{P}(D < g\,|\,X_1)}\\ &= \exp(\alpha_g + \beta_1 X_1)\end{aligned}$$

We will first consider the special case where the exposure is the only independent variable and is coded 1 and 0. Recall that the odds comparing $D \geq g$ vs. $D < g$ is e to the α_g plus β_1 times X_1. To assess the effect of the exposure on the outcome, we formulate the ratio of the odds of $D \geq g$ for comparing $X_1 = 1$ and $X_1 = 0$ (i.e., the odds ratio for $X_1 = 1$ vs. $X_1 = 0$).

$$\text{OR} = \frac{\text{P}(D \geq g \mid X_1 = 1)/\text{P}(D < g \mid X_1 = 1)}{\text{P}(D \geq g \mid X_1 = 0)/\text{P}(D < g \mid X_1 = 0)}$$

$$= \frac{\exp[\alpha_g + \beta_1(1)]}{\exp[\alpha_g + \beta_1(0)]} = \frac{\exp(\alpha_g + \beta_1)}{\exp(\alpha_g)}$$

$$= \text{e}^{\beta_1}$$

This is calculated, as shown on the left, as the odds that the disease outcome is greater than or equal to g if X_1 equals 1, divided by the odds that the disease outcome is greater than or equal to g if X_1 equals 0.

Substituting the expression for the odds in terms of the regression parameters, the odds ratio for $X_1 = 1$ vs. $X_1 = 0$ in the comparison of disease levels $\geq g$ to levels $<$ g is then e to the β_1.

General case
(levels X_1^{**} and X_1^{*} of X_1)

$$\text{OR} = \frac{\exp(\alpha_g + \beta_1 X_1^{**})}{\exp(\alpha_g + \beta_1 X_1^{*})}$$

$$= \frac{\exp(\alpha_g)\ \exp(\beta_1 X_1^{**})}{\exp(\alpha_g)\ \exp(\beta_1 X_1^{*})}$$

$$= \exp\left[\beta_1(X_1^{**} - X_1^{*})\right]$$

To compare any two levels of the exposure variable, X_1^{**} and X_1^{*}, the odds ratio formula is e to the β_1 times the quantity X_1^{**} minus X_1^{*}.

CIs: same method as SLR to compute CIs

General case (levels X_1^{**} and X_1^{*} of X_1)

95% CI:

$$\exp\left[\hat{\beta}_1(X_1^{**} - X_1^{*}) \pm 1.96(X_1^{**} - X_1^{*})s_{\hat{\beta}_1}\right]$$

Confidence interval estimation is also analogous to standard logistic regression. The general large-sample formula for a 95% confidence interval, for any two levels of the independent variable (X_1^{**} and X_1^{*}), is shown on the left.

EXAMPLE

Black/White Cancer Survival Study

Test of proportional odds assumption:
H_0: assumption holds
Score statistic: $\chi^2 = 0.0008$, df $= 1$,
$P = 0.9779$.
Conclusion: fail to reject null

Returning to our tumor-grade example, the results for the model examining tumor grade and RACE are presented next. The results were obtained from running PROC LOGISTIC in SAS (see Appendix).

We first check the proportional odds assumption with a *Score test*. The test statistic, with one degree of freedom for the one odds ratio parameter being tested, was clearly not significant, with a P-value of 0.9779. We therefore fail to reject the null hypothesis (i.e., that the assumption holds) and can proceed to examine the model output.

EXAMPLE (*continued*)

Variable	Estimate	S.E.
Intercept 1 ($\hat{\alpha}_2$)	−1.7388	0.1765
Intercept 2 ($\hat{\alpha}_1$)	−0.0089	0.1368
RACE	0.7555	0.2466

$\widehat{\text{OR}} = \exp(0.7555) = 2.13$

With this ordinal model, there are two intercepts, one for each comparison, but there is only one estimated beta for the effect of RACE. The odds ratio for RACE is e to β_1. In our example, the odds ratio equals exp(0.7555) or 2.13. [Note: SAS's LOGISTIC procedure was used with a "descending" option so that Intercept 1 compares D $\geq$ 2 to D $<$ 2, whereas Intercept 2 compares D $\geq$ 1 to D $<$ 1].

Interpretation of OR

Black vs. white women with endometrial cancer over twice as likely to have more severe tumor grade:

Since $\widehat{\text{OR}}$ $(D \geq 2) = \widehat{\text{OR}}$ $(D \geq 1) = 2.13$

The results indicate that for this sample of women with invasive endometrial cancer, black women were over twice (i.e., 2.13) as likely as white women to have tumors that were categorized as poorly differentiated vs. moderately differentiated or well differentiated *and* over twice as likely as white women to have tumors classified as poorly differentiated or moderately differentiated vs. well differentiated. To summarize, in this cohort, black women were over twice as likely to have a more severe grade of endometrial cancer compared with white women.

Interpretation of intercepts (α_g)

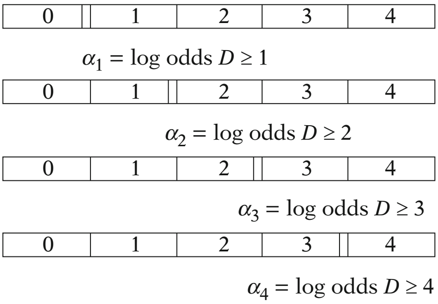
α_g = log odds of $D \geq g$ where all independent variables equal zero;

$g = 1, 2, 3, \ldots, G - 1$

$$\alpha_g > \alpha_{g+1}$$
$$\Downarrow$$
$$\alpha_1 > \alpha_2 > \cdots > \alpha_{G-1}$$

What is the interpretation of the intercept? The intercept α_g is the log odds of $D \geq g$ where all the independent variables are equal to zero. This is similar to the interpretation of the intercept for other logistic models except that, with the proportional odds model, we are modeling the log odds of several inequalities. This yields several intercepts, with each intercept corresponding to the log odds of a different inequality (depending on the value of g). Moreover, the log odds of $D \geq g$ is greater than the log odds of $D \geq (g + 1)$ (assuming category g is nonzero). This means that $\alpha_1 > \alpha_2 \cdots > \alpha_{G-1}$.

Illustration


0
1
2
3
4
α_1 = log odds $D \geq 1$
α_2 = log odds $D \geq 2$
α_3 = log odds $D \geq 3$
α_4 = log odds $D \geq 4$

As the picture on the left illustrates, with five categories ($D = 0, 1, 2, 3, 4$), the log odds of $D \geq 1$ is greater than the log odds of $D \geq 2$, since for $D \geq 1$, the outcome can be in categories 1, 2, 3, or 4, whereas for $D \geq 2$, the outcome can only be in categories 2, 3, or 4. Thus, there is one more outcome category (category 1) contained in the first inequality. Similarly, the log odds of $D \geq 2$ is greater than the log odds of $D \geq 3$, and the log odds of $D \geq 3$ is greater than the log odds of $D \geq 4$.

EXAMPLE (*continued*)

95% confidence interval for OR

$$95\%\ \text{CI} = \exp[0.7555 \pm 1.96\ (0.2466)] = (1.31, 3.45)$$

Returning to our example, the 95% confidence interval for the OR for AGE is calculated as shown on the left.

Hypothesis testing

Likelihood ratio test or Wald test
H_0: $\beta_1 = 0$

Hypothesis testing about parameter estimates can be done using either the likelihood ratio test or the Wald test. The null hypothesis is that β_1 is equal to 0.

Wald test

$$Z = \frac{0.7555}{0.2466} = 3.06, \quad P = 0.002$$

In the tumor grade example, the P-value for the Wald test of the beta coefficient for RACE is 0.002, indicating that RACE is significantly associated with tumor grade at the 0.05 level.

IV. Extending the Ordinal Model

$$\mathrm{P}(D \geq \mathrm{g}\,|\,\mathbf{X}) = \frac{1}{1+\exp[-(\alpha_g + \sum_{i=1}^{k} \beta_i X_i)]},$$

where $g = 1, 2, 3, \ldots, G{-}1$

Note: $\mathrm{P}(D \geq 0\,|\,\mathbf{X}) = 1$

Expanding the model to add more independent variables is straightforward. The model with k independent variables is shown on the left.

$$\text{odds} = \frac{\mathrm{P}(D \geq g\,|\,\mathbf{X})}{\mathrm{P}(D < g\,|\,\mathbf{X})}$$
$$= \exp(\alpha_g + \sum_{i=1}^{k} \beta_i X_j)$$

The *odds* for the outcome greater than or equal to level g is then e to the quantity α_g plus the summation the X_i for each of the k independent variable times its beta.

$\mathrm{OR} = \exp(\beta_{\mathrm{i}})$, if X_i is coded (0, 1)

The odds ratio is calculated in the usual manner as e to the β_i, if X_i is coded 0 or 1. As in standard logistic regression, the use of multiple independent variables allows for the estimation of an odds ratio for one variable controlling for the effects of the other covariates in the model.

EXAMPLE

$$D = \text{GRADE} = \begin{cases} 0 & \text{if well differentiated} \\ 1 & \text{if moderately differentiated} \\ 2 & \text{if poorly differentiated} \end{cases}$$

$$X_1 = \text{RACE} = \begin{cases} 0 & \text{if white} \\ 1 & \text{if black} \end{cases}$$

$$X_2 = \text{ESTROGEN} = \begin{cases} 0 & \text{if never user} \\ 1 & \text{if ever user} \end{cases}$$

$$\mathrm{P}(D \geq g\,|\,\mathbf{X}) = \frac{1}{1+\exp[-(\alpha_g + \beta_1 X_1 + \beta_2 X_2)]},$$

where $X_1 = \text{RACE } (0, 1)$
$X_2 = \text{ESTROGEN } (0, 1)$
$g = 1, 2$

To illustrate, we return to our endometrial tumor grade example. Suppose we wish to consider the effects of estrogen use as well as RACE on GRADE. ESTROGEN is coded as 1 for ever user and 0 for never user.

The model now contains two predictor variables: X_1 = RACE and X_2 = ESTROGEN.

EXAMPLE (*continued*)

$$\text{odds} = \frac{P(D \geq 2 \mid \mathbf{X})}{P(D < 2 \mid \mathbf{X})} = \exp(\alpha_2 + \beta_1 X_1 + \beta_2 X_2)$$

different αs

same βs

$$\text{odds} = \frac{P(D \geq 1 \mid \mathbf{X})}{P(D < 1 \mid \mathbf{X})} = \exp(\alpha_1 + \beta_1 X_1 + \beta_2 X_2)$$

The odds that the tumor grade is in a category greater than or equal to category 2 (i.e., poorly differentiated) vs. in categories less than 2 (i.e., moderately or well differentiated) is e to the quantity α_2 plus the sum of $\beta_1 X_1$ plus $\beta_2 X_2$.

Similarly, the odds that the tumor grade is in a category greater than or equal to category 1 (i.e., moderately or poorly differentiated) vs. in categories less than 1 (i.e., well differentiated) is e to the quantity α_1 plus the sum of $\beta_1 X_1$ plus $\beta_2 X_2$. Although the alphas are different, the betas are the same.

Test of proportional odds assumption

H_0: assumption holds
Score statistic: $\chi^2 = 0.9051$, 2 df,
$P = 0.64$
Conclusion: fail to reject null

Before examining the model output, we first check the proportional odds assumption with a Score test. The test statistic has two degrees of freedom because we have two fewer parameters in the ordinal model compared to the corresponding polytomous model. The results are not statistically significant, with a *P*-value of 0.64. We therefore fail to reject the null hypothesis that the assumption holds and can proceed to examine the remainder of the model results.

Variable	Estimate	S.E.	Symbol
Intercept 1	−1.2744	0.2286	$\hat{\alpha}_2$
Intercept 2	0.5107	0.2147	$\hat{\alpha}_1$
RACE	0.4270	0.2720	$\hat{\beta}_1$
ESTROGEN	−0.7763	0.2493	$\hat{\beta}_2$

The output for the analysis is shown on the left. There is only one beta estimate for each of the two predictor variables in the model. Thus, there are a total of four parameters in the model, including the two intercepts.

EXAMPLE (*continued*)

Odds ratio

$$\widehat{\text{OR}} = \exp \hat{\beta}_1 = \exp(0.4270) = 1.53$$

The estimated odds ratio for the effect of RACE, controlling for the effect of ESTROGEN, is e to the $\hat{\beta}_1$, which equals e to the 0.4270 or 1.53.

95% confidence interval

$$95\%\ \text{CI} = \exp[0.4270 \pm 1.96(0.2720)] = (0.90, 2.61)$$

The 95% confidence interval for the odds ratio is e to the quantity $\hat{\beta}_1$ plus or minus 1.96 times the estimated standard error of the beta coefficient for RACE. In our two-predictor example, the standard error for RACE is 0.2720 and the 95% confidence interval is calculated as 0.90 to 2.61. The confidence interval contains one, the null value.

Wald test

$$H_0 : \beta_1 = 0$$

$$Z = \frac{0.4270}{0.2720} = 1.57, \quad P = 0.12$$

Conclusion: fail to reject H_0

If we perform the Wald test for the significance of $\hat{\beta}_1$, we find that it is not statistically significant in this two-predictor model ($P = 0.12$). The addition of ESTROGEN to the model has resulted in a decrease in the estimated effect of RACE on tumor grade, suggesting that failure to control for ESTROGEN biases the effect of RACE away from the null.

V. Likelihood Function for Ordinal Model

$$\text{odds} = \frac{P}{1-P}$$

so solving for P,

$$P = \frac{\text{odds}}{\text{odds} + 1} = \frac{1}{1 + \left(\frac{1}{\text{odds}}\right)}$$

Next, we briefly discuss the development of the likelihood function for the proportional odds model. To formulate the likelihood, we need the probability of the observed outcome for each subject. An expression for these probabilities in terms of the model parameters can be obtained from the relationship $P = \text{odds}/(\text{odds} + 1)$, or the equivalent expression $P = 1/[1 + (1/\text{odds})]$.

$$\mathrm{P}(D = g) = [\mathrm{P}(D \geq g)] - [\mathrm{P}(D \geq g + 1)\ ]$$

For $g = 2$
$\mathrm{P}(D = 2) = \mathrm{P}(D \geq 2) - P(D \geq 3)$

Use relationship to obtain probability that individual is in given outcome category.

In the proportional odds model, we model the probability of $D \geq g$. To obtain an expression for the probability of $D = g$, we can use the relationship that the probability $(D = g)$ is equal to the probability of $D \geq g$ minus the probability of $D \geq (g + 1)$. For example, the probability that D equals 2 is equal to the probability that D is greater than or equal to 2 minus the probability that D is greater than or equal to 3. In this way we can use the model to obtain an expression for the probability that an individual is in a specific outcome category for a given pattern of covariates (**X**).

L is product of individual contributions.

$$\prod_{j=1}^{n} \prod_{g=0}^{G-1} \mathrm{P}(D = g \,|\, \mathbf{X})^{y_{jg}},$$

where

$$y_{jg} = \begin{cases} 1 & \text{if the } j\text{th subject has } D = g \\ 0 & \text{if otherwise} \end{cases}$$

The likelihood (L) is then calculated in the same manner discussed previously in the section on polytomous regression – that is, by taking the product of the individual contributions.

VI. Ordinal vs. Multiple Standard Logistic Regressions

Proportional odds model: order of outcome considered.

Alternative: several logistic regression models

The proportional odds model takes into account the effect of an exposure on an ordered outcome and yields one odds ratio summarizing that effect across outcome levels. An alternative approach is to conduct a series of logistic regressions with different dichotomized outcome variables. A separate odds ratio for the effect of the exposure can be obtained for each of the logistic models.

Original variable: 0, 1, 2, 3
Recoded:
≥ 1 vs. < 1, ≥ 2 vs. < 2 , and ≥ 3 vs. < 3

For example, in a four-level outcome variable, coded as 0, 1, 2, and 3, we can define three new outcomes: greater than or equal to 1 vs. less than 1, greater than or equal to 2 vs. less than 2, and greater than or equal to 3 vs. less than 3.

Three separate logistic regressions

Three sets of parameters

$\alpha \geq 1$ vs. < 1, $\beta \geq 1$ vs. < 1

$\alpha \geq 2$ vs. < 2, $\beta \geq 2$ vs. < 2

$\alpha \geq 3$ vs. < 3, $\beta \geq 3$ vs. < 3

With these three dichotomous outcomes, we can perform three separate logistic regressions. In total, these three regressions would yield three intercepts and three estimated beta coefficients for each independent variable in the model.

Logistic models (three parameters)	Proportional odds model (one parameter)
$\beta \geq 1$ vs. < 1	
$\beta \geq 2$ vs. < 2	β
$\beta \geq 3$ vs. < 3	

If the proportional odds assumption is reasonable, then using the proportional odds model allows us to summarize the relationship between the outcome and each independent variable with one parameter instead of three.

Is the proportional odds assumption met?

- Crude ORs "close"?
  (No control of confounding)

The key question is whether or not the proportional odds assumption is met. There are several approaches to checking the assumption. Calculating and comparing the crude odds ratios is the simplest method, but this does not control for confounding by other variables in the model.

- Beta coefficients in separate logistic models similar?
  (Not a statistical test)

Is $\beta_{\geq 1 \text{ vs. } <1} \cong \beta_{\geq 2 \text{ vs. } <2} \cong \beta_{\geq 3 \text{ vs. } <3}$?

Running the separate (e.g., 3) logistic regressions allows the investigator to compare the corresponding odds ratio parameters for each model and assess the reasonableness of the proportional odds assumption in the presence of possible confounding variables. Comparing odds ratios in this manner is not a substitute for a statistical test, although it does provide the means to compare parameter estimates. For the four-level example, we would check whether the three coefficients for each independent variable are similar to each other.

- Score test provides a test of proportional odds assumption

H_0: assumption holds

The Score test enables the investigator to perform a statistical test on the proportional odds assumption. With this test, the null hypothesis is that the proportional odds assumption holds. However, failure to reject the null hypothesis does not necessarily mean the proportional odds assumption is reasonable. It could be that there are not enough data to provide the statistical evidence to reject the null.

If assumption not met, may

- Use polytomous logistic model
- Use different ordinal model
- Use separate logistic models

If the assumption does not appear to hold, one option for the researcher would be to use a polytomous logistic model. Another alternative would be to select an ordinal model other than the proportional odds model. A third option would be to use separate logistic models. The approach selected should depend on whether the assumptions underlying the specific model are met and on the type of inferences the investigator wishes to make.

VII. SUMMARY

✓ Chapter 13: Ordinal Logistic Regression

This presentation is now complete. We have described a method of analysis, ordinal regression, for the situation where the outcome variable has more than two ordered categories. The proportional odds model was described in detail. This may be used if the proportional odds assumption is reasonable.

We suggest that you review the material covered here by reading the detailed outline that follows. Then do the practice exercises and test.

Chapter 14: Logistic Regression for Correlated Data: GEE

All of the models presented thus far have assumed that observations are statistically independent, (i.e., are not correlated). In the next chapter (Chap. 14), we consider one approach for dealing with the situation in which study outcomes are not independent.

Detailed Outline

I. Overview (page 466)

A. Focus: modeling outcomes with more than two levels.

B. Ordinal outcome variables.

II. Ordinal logistic regression: The proportional odds model (pages 466–472)

A. Ordinal outcome: variable categories have a natural order.

B. Proportional odds assumption: the odds ratio is invariant to where the outcome categories are dichotomized.

C. The form for the proportional odds model with one independent variable (X_1) for an outcome (D) with G levels ($D = 0, 1, 2, \ldots, G-1$) is

$$\mathrm{P}(D \geq g \,|\, X_1) = \frac{1}{1 + \exp[-(\alpha_g + \beta_1 X_1)]},$$

where $g = 1, 2, \ldots, G - 1$

III. Odds ratios and confidence limits (pages 472–475)

A. Computation of the OR in ordinal regression is analogous to standard logistic regression, except that there is a single odds ratio for all comparisons.

B. The general formula for the odds ratio for any two levels of the predictor variable (X_1^{**} and X_1^{*}) is

$$\mathrm{OR} = \exp[\beta_1(X_1^{**} - X_1^{*})]$$

for a model with one independent variable (X_1).

C. Confidence interval estimation is analogous to standard logistic regression.

D. The general large-sample formula for a 95% confidence interval for any two levels of the independent variable (X_1^{**} and X_1^{*}) is

$$\exp\left[\hat{\beta}_1(X_1^{**} - X_1^{*}) \pm 1.96(X_1^{**} - X_1^{*})s_{\hat{\beta}_1}\right]$$

E. The likelihood ratio test is used to test hypotheses about the significance of the predictor variable(s).

i. There is one estimated coefficient for each predictor.

ii. The null hypothesis is that the beta coefficient (for a given predictor) is equal to zero.

iii. The test compares the log likelihood of the full model with the predictor(s) to that of the reduced model without the predictor(s).

F. The Wald test is analogous to standard logistic regression.

IV. Extending the ordinal model (pages 476–478)

A. The general form of the proportional odds model for G outcome categories and k independent variables is

$$P(D \geq g \,|\, \mathbf{X}) = \frac{1}{1 + \exp\left[-(\alpha_g + \sum_{i=1}^{k} \beta_i X_i)\right]}$$

B. The calculation of the odds ratio, confidence intervals, and hypothesis testing using the likelihood ratio and Wald tests remain the same.

C. Interaction terms can be added and tested in a manner analogous to standard logistic regression.

V. Likelihood function for ordinal model (pages 478–479)

A. For an outcome variable with G categories, the likelihood function is

$$\prod_{j=1}^{n} \prod_{g=0}^{G-1} P(D = g \,|\, \mathbf{X}^{y_{jg}}),$$

where

$$y_{jg} = \begin{cases} 1 & \text{if the jth subject has } D = g \\ 0 & \text{if otherwise} \end{cases}$$

where n is the total number of subjects, $g = 0, 1, \ldots, G-1$ and
$P(D = g|\mathbf{X}) = [P(D \geq g|\mathbf{X})] - [P(D \geq g + 1)\,|\,\mathbf{X})]$.

VI. Ordinal vs. multiple standard logistic regressions (pages 479–481)

A. Proportional odds model: order of outcome considered.

B. Alternative: several logistic regressions models

i. One for each cut-point dichotomizing the outcome categories.

ii. Example: for an outcome with four categories (0, 1, 2, 3), we have three possible models.

C. If the proportional odds assumption is met, it allows the use of one parameter estimate for the effect of the predictor, rather than separate estimates from several standard logistic models.

D. To check if the proportional odds assumption is met:
   i. Evaluate whether the crude odds ratios are "close".
   ii. Evaluate whether the odds ratios from the standard logistic models are similar:
      a. Provides control of confounding but is not a statistical test.
   iii. Perform a Score test of the proportional odds assumption.

E. If assumption is not met, can use a polytomous model, consider use of a different ordinal model, or use separate logistic regressions.

VII. Summary (page 481)

Practice Exercises

Suppose we are interested in assessing the association between tuberculosis and degree of viral suppression in HIV-infected individuals on antiretroviral therapy, who have been followed for 3 years in a hypothetical cohort study. The outcome, tuberculosis, is coded as none ($D = 0$), latent ($D = 1$), or active ($D = 2$). Degree of viral suppression (VIRUS) is coded as undetectable (VIRUS = 0) or detectable (VIRUS = 1). Previous literature has shown that it is important to consider whether the individual has progressed to AIDS (no = 0, yes = 1) and is compliant with therapy (COMPLIANCE: no = 1, yes = 0). In addition, AGE (continuous) and GENDER (female = 0, male = 1) are potential confounders. Also there may be interaction between progression to AIDS and COMPLIANCE with therapy (AIDSCOMP = AIDS × COMPLIANCE).

We decide to run a proportional odds logistic regression to analyze these data. Output from the ordinal regression is shown below. (The results are hypothetical.) The descending option was used, so Intercept 1 pertains to the comparison $D \geq 2$ to $D < 2$ and Intercept 2 pertains to the comparison $D \geq 1$ to $D < 1$.

Variable	Coefficient	S.E.
Intercept 1 (α_2)	−2.98	0.20
Intercept 2 (α_1)	−1.65	0.18
VIRUS	1.13	0.09
AIDS	0.82	0.08
COMPLIANCE	0.38	0.14
AGE	0.04	0.03
GENDER	0.35	0.19
AIDSCOMP	0.31	0.14

1. State the form of the ordinal model in terms of variables and unknown parameters.
2. For the above model, state the fitted model in terms of variables and estimated coefficients.
3. Compute the estimated odds ratio for a 25-year-old noncompliant male with a detectable viral load, who has progressed to AIDS, compared with a similar female. Consider the outcome comparison active or latent tuberculosis versus none ($D \geq 1$ vs. $D < 1$).
4. Compute the estimated odds ratio for a 38-year-old noncompliant male with a detectable viral load, who has progressed to AIDS, compared with a similar female. Consider the outcome comparison active tuberculosis versus latent or none ($D \geq 2$ vs. $D < 2$).
5. Estimate the odds of a compliant 20-year-old female, with an undetectable viral load and who has not progressed to AIDS, of having active tuberculosis ($D \geq 2$).
6. Estimate the odds of a compliant 20-year-old female, with an undetectable viral load and who has not progressed to AIDS, of having latent or active tuberculosis ($D \geq 1$).
7. Estimate the odds of a compliant 20-year-old male, with an undetectable viral load and who has not progressed to AIDS, of having latent or active tuberculosis ($D \geq 1$).
8. Estimate the odds ratio for noncompliance vs. compliance. Consider the outcome comparison active tuberculosis vs. latent or no tuberculosis ($D \geq 2$ vs. $D < 2$).

Test

True or False (Circle T or F)

T F 1. The disease categories absent, mild, moderate, and severe can be ordinal.

T F 2. In an ordinal logistic regression (using a proportional odds model) in which the outcome variable has five levels, there will be four intercepts.

T F 3. In an ordinal logistic regression in which the outcome variable has five levels, each independent variable will have four estimated coefficients.

T F 4. If the outcome D has seven levels (coded 1, 2, . . . , 7), then $\mathrm{P}(D \geq 4)/\mathrm{P}(D < 4)$ is an example of an odds.

T F 5. If the outcome D has seven levels (coded 1, 2, . . . , 7), an assumption of the proportional odds model is that $\mathrm{P}(D \geq 3)/\mathrm{P}(D < 3)$ is assumed equal to $\mathrm{P}(D \geq 5)/\mathrm{P}(D < 5)$.

T F 6. If the outcome D has seven levels (coded 1, 2, . . . , 7) and an exposure E has two levels (coded 0 and 1), then an assumption of the proportional odds model is that $[\mathrm{P}(D \geq 3|E = 1)/\mathrm{P}(D < 3|E = 1)]/[\mathrm{P}(D \geq 3|E = 0)/\mathrm{P}(D < 3|E = 0)]$ is assumed equal to $[\mathrm{P}(D \geq 5|E = 1)/\mathrm{P}(D < 5|E = 1)]/[\mathrm{P}(D \geq 5|E = 0)/\mathrm{P}(D < 5|E = 0)]$.

T F 7. If the outcome D has four categories coded $D = 0, 1, 2, 3$, then the log odds of $D \geq 2$ is greater than the log odds of $D \geq 1$.

T F 8. Suppose a four level outcome D coded $D = 0, 1, 2, 3$ is recoded $D^* = 1, 2, 7, 29$, then the choice of using D or D^* as the outcome in a proportional odds model has no effect on the parameter estimates as long as the order in the outcome is preserved.

9. Suppose the following proportional odds model is specified assessing the effects of AGE (continuous), GENDER (female = 0, male = 1), SMOKE (nonsmoker = 0, smoker = 1), and hypertension status (HPT) (no = 0, yes = 1) on four progressive stages of disease ($D = 0$ for absent, $D = 1$ for mild, $D = 2$ for severe, and $D = 3$ for critical).

$$\ln\frac{\mathrm{P}(D \geq g\,|\,\mathbf{X})}{\mathrm{P}(D < g\,|\,\mathbf{X})} = \alpha_g + \beta_1\mathrm{AGE} + \beta_2\mathrm{GENDER} + \beta_3\mathrm{SMOKE} + \beta_4\mathrm{HPT},$$

where $g = 1, 2, 3$.

Use the model to obtain an expression for the odds of a severe or critical outcome ($\mathrm{D} \geq 2$) for a 40-year-old male smoker without hypertension.

10. Use the model in Question 9 to obtain the odds ratio for the mild, severe, or critical stage of disease (i.e., $D \geq 1$)] comparing hypertensive smokers vs. nonhypertensive nonsmokers, controlling for AGE and GENDER.
11. Use the model in Question 9 to obtain the odds ratio for critical disease only ($D \geq 3$) comparing hypertensive smokers vs. nonhypertensive nonsmokers, controlling for AGE and GENDER. Compare this odds ratio to that obtained for Question 10.
12. Use the model in Question 9 to obtain the odds ratio for mild or no disease ($D < 2$) comparing hypertensive smokers vs. nonhypertensive nonsmokers, controlling for AGE and GENDER.

Answers to Practice Exercises

1. Ordinal model

$$\ln\left[\frac{\mathrm{P}(D \geq g\,|\,\mathbf{X})}{\mathrm{P}(D < g\,|\,\mathbf{X})}\right] = \alpha_g + \beta_1\mathrm{VIRUS} + \beta_2\mathrm{AIDS} + \beta_3\mathrm{COMPLIANCE} + \beta_4\mathrm{AGE} + \beta_5\mathrm{GENDER} + \beta_6\mathrm{AIDSCOMP},$$

where $g = 1, 2$

2. Ordinal fitted model

$$\hat{\ln}\left[\frac{\mathrm{P}(D \geq 2\,|\,\mathbf{X})}{\mathrm{P}(D < 2\,|\,\mathbf{X})}\right] = -2.98 + 1.13\mathrm{VIRUS} + 0.82\mathrm{AIDS} + 0.38\mathrm{COMPLIANCE} + 0.04\mathrm{AGE} + 0.35\mathrm{GENDER} + 0.31\mathrm{AIDSCOMP},$$

$$\hat{\ln}\left[\frac{\mathrm{P}(D \geq 1\,|\,\mathbf{X})}{\mathrm{P}(D < 1\,|\,\mathbf{X})}\right] = -1.65 + 1.13\mathrm{VIRUS} + 0.82\mathrm{AIDS} + 0.38\mathrm{COMPLIANCE} + 0.04\mathrm{AGE} + 0.35\mathrm{GENDER} + 0.31\mathrm{AIDSCOMP}.$$

3. $\widehat{\mathrm{OR}} = \exp(0.35) = 1.42$
4. $\widehat{\mathrm{OR}} = \exp(0.35) = 1.42$
5. Estimated odds $= \exp[-2.98 + 20(0.04)] = 0.11$
6. Estimated odds $= \exp[-1.65 + 20(0.04)] = 0.43$
7. Estimated odds $= \exp[-1.65 + 20(0.04) + 0.35] = 0.61$
8. Estimated odds ratios for noncompliant (COMPLIANCE = 1) vs. compliant (COMPLIANCE = 0) subjects:

   For AIDS = 0: $\exp(0.38) = 1.46$

   For AIDS = 1: $\exp(0.38 + 0.31) = 1.99$

14 Logistic Regression for Correlated Data: GEE

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health, DOI 10.1007/978-1-4419-1742-3_14, © Springer Science+Business Media, LLC 2010

Introduction

In this chapter, the logistic model is extended to handle outcome variables that have dichotomous correlated responses. The analytic approach presented for modeling this type of data is the generalized estimating equations (GEE) model, which takes into account the correlated nature of the responses. If such correlations are ignored in the modeling process, then incorrect inferences may result.

The form of the GEE model and its interpretation are developed. A variety of correlation structures that are used in the formulation of the model are described. An overview of the mathematical foundation for the GEE approach is also presented, including discussions of generalized linear models, score equations, and "score-like" equations. In the next chapter (Chap. 12), examples are presented to illustrate the application and interpretation of GEE models. The final chapter in the text (Chap. 13) describes alternate approaches for the analysis of correlated data.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

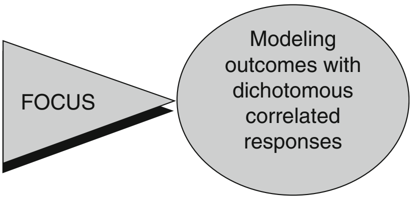
Objectives

Upon completing this chapter, the learner should be able to:

1. State or recognize examples of correlated responses.
2. State or recognize when the use of correlated analysis techniques may be appropriate.
3. State or recognize an appropriate data layout for a correlated analysis.
4. State or recognize the form of a GEE model.
5. State or recognize examples of different correlation structures that may be used in a GEE model.

Presentation

I. Overview



In this chapter, we provide an introduction to modeling techniques for use with dichotomous outcomes in which the responses are correlated. We focus on one of the most commonly used modeling techniques for this type of analysis, known as generalized estimating equations or GEE, and we describe how the GEE approach is used to carry out logistic regression for correlated dichotomous responses.

Examples of correlated responses:

1. Different members of the same house-hold.
2. Each eye of the same person.
3. Several bypass grafts on the same subject.
4. Monthly measurements on the same subject.

For the modeling techniques discussed previously, we have made an assumption that the responses are independent. In many research scenarios, this is not a reasonable assumption. Examples of correlated responses include (1) observations on different members of the same household, (2) observations on each eye of the same person, (3) results (e.g., success/failure) of several bypass grafts on the same subject, and (4) measurements repeated each month over the course of a year on the same subject. The last is an example of a longitudinal study, since individuals' responses are measured repeatedly over time.

Observations can be grouped into clusters:

Example No.	Cluster	Source of observation
1	Household	Household members
2	Subject	Eyes
3	Subject	Bypass grafts
4	Subject	Monthly repeats

For the above-mentioned examples, the observations can be grouped into clusters. In example 1, the clusters are households, whereas the observations are the individual members of each household. In example 4, the clusters are individual subjects, whereas the observations are the monthly measurements taken on the subject.

Assumption:

$$\text{Responses} \begin{cases} \text{correlated within clusters} \\ \text{independent between clusters} \end{cases}$$

A common assumption for correlated analyses is that the responses are correlated within the same cluster but are independent between different clusters.

.

Ignoring within-cluster correlation

$\Downarrow$

Incorrect inferences

In analyses of correlated data, the correlations between subject responses often are ignored in the modeling process. An analysis that ignores the correlation structure may lead to incorrect inferences.

II. An Example (Infant Care Study)

GEE vs. standard logistic regression (ignores correlation)

- Statistical inferences may differ
- Similar use of output

We begin by illustrating how statistical inferences may differ depending on the type of analysis performed. We shall compare a generalized estimating equations (GEE) approach with a standard logistic regression that ignores the correlation structure. We also show the similarities of these approaches in utilizing the output to obtain and interpret odds ratio estimates, their corresponding confidence intervals, and tests of significance.

Data source: Infant Care Study in Brazil

Subjects: 168 infants
136 with complete data

The data were obtained from an infant care health intervention study in Brazil (Cannon et al., 2001). As a part of that study, height and weight measurements were taken each month from 168 infants over a 9-month period. Data from 136 infants with complete data on the independent variables of interest are used for this example.

Response (D): weight-for-height standardized (z) score

$$D = \begin{cases} 1 & \text{if } z < -1 \text{ (``Wasting'')} \\ 0 & \text{otherwise} \end{cases}$$

Independent variables:
BIRTHWGT (in grams)
GENDER

$$\text{DIARRHEA} = \begin{cases} 1 & \text{if symptoms present in past month} \\ 0 & \text{otherwise} \end{cases}$$

The response (D) is derived from a weight-for-height standardized score (i.e., z-score) based on the weight-for-height distribution of a reference population. A weight-for-height measure of more than one standard deviation below the mean (i.e., $z < -1$) indicates "wasting". The dichotomous outcome for this study is coded 1 if the z-score is less than negative 1 and 0 otherwise. The independent variables are BIRTHWGT (the weight in grams at birth), GENDER, and DIARRHEA (a dichotomous variable indicating whether the infant had symptoms of diarrhea that month).

Infant Care Study: Sample Data

From three infants: five (of nine) observations listed for each

```
IDNO    MO   OUTCOME   BIRTHWGT   GENDER   DIARRHEA
00282   1      0        2000       Male       0
00282   2      0        2000       Male       0
00282   3      1        2000       Male       1
  .     .      .          .         .
00282   8      0        2000       Male       1
00282   9      0        2000       Male       0
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
00283   1      0        2950      Female      0
00283   2      0        2950      Female      0
00283   3      1        2950      Female      0
  .     .      .          .         .
00283   8      0        2950      Female      0
00283   9      0        2950      Female      0
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
00287   1      1        3250       Male       1
00287   2      1        3250       Male       1
00287   3      0        3250       Male       0
  .     .      .          .         .
00287   8      0        3250       Male       0
00287   9      0        3250       Male       0
. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
```

IDNO: identification number

MO: observation month (provides order to subject-specific measurements)

OUTCOME: dichotomized z-score (values can change month to month)

On the left, we present data on three infants to illustrate the layout for correlated data. Five of nine monthly observations are listed per infant. In the complete data on 136 infants, each child had at least 5 months of observations, and 126 (92.6%) had complete data for all 9 months.

The variable IDNO is the number that identifies each infant. The variable MO indicates which month the outcome measurement was taken. This variable is used to provide order for the data within a cluster. Not all clustered data have an inherent order to the observations within a cluster; however, in longitudinal studies such as this, specific measurements are ordered over time.

The variable OUTCOME is the dichotomized weight-for-height z-score indicating the presence or absence of wasting. Notice that the outcome can change values from month to month within a cluster.

Independent variables:

1. Time-dependent variable: can vary month to month within a cluster
   DIARRHEA: dichotomized variable for presence of symptoms
2. Time-independent variables: do not vary month to month within a cluster
   BIRTHWGT
   GENDER

The independent variable DIARRHEA can also change values month to month. If symptoms of diarrhea are present in a given month, then the variable is coded 1; otherwise it is coded 0. DIARRHEA is thus a *time-dependent* variable. This contrasts with the variables BIRTHWGT and GENDER, which do not vary within a cluster (i.e., do not change month to month). BIRTHWGT and GENDER are *time-independent* variables.

In general, with longitudinal data, independent variables may be either

1. Time-dependent
   or
2. Time-independent

Outcome variable generally varies within a cluster

In general, with longitudinal data, independent variables may or may not vary within a cluster. A time-dependent variable can vary in value, whereas a time-independent variable does not. The values of the *outcome* variable, in general, will vary within a cluster.

Goal of analysis: to account for outcome variation within and between clusters

A correlated analysis attempts to account for the variation of the outcome from both within and between clusters.

Model for Infant Care Study:

$$\begin{aligned}\text{logit P}(D=1\,|\,\mathbf{X}) = \beta_0 &+ \beta_1\text{BIRTHWGT}\\ &+ \beta_2\text{GENDER}\\ &+ \beta_3\text{DIARRHEA}\end{aligned}$$

We state the model for the Infant Care Study example in logit form as shown on the left. In this chapter, we use the notation β_0 to represent the intercept rather than α, as α is commonly used to represent the correlation parameters in a GEE model.

GEE Model (GENMOD output)

```
                          Empirical      Wald
Variable   Coefficient     Std Err     p-value

INTERCEPT    −1.3978        1.1960      0.2425
BIRTHWGT     −0.0005        0.0003      0.1080
GENDER        0.0024        0.5546      0.9965
DIARRHEA      0.2214        0.8558      0.7958
```

Interpretation of GEE model similar to SLR

$$\left.\begin{array}{l}\text{OR estimates}\\ \text{Confidence intervals}\\ \text{Wald test statistics}\end{array}\right\}\begin{array}{l}\text{Use same}\\ \text{formulas}\end{array}$$

$$\left.\begin{array}{l}\text{Underlying assumptions}\\ \text{Method of parameter}\\ \quad\text{estimation}\end{array}\right\}\text{Differ}$$

Next, the output obtained from running a GEE model using the GENMOD procedure in SAS is presented. This model accounts for the correlations among the monthly outcome within each of the 136 infant clusters. Odds ratio estimates, confidence intervals, and Wald test statistics are obtained using the GEE model output in the same manner (i.e., with the same formulas) as we have shown previously using output generated from running a standard logistic regression. The interpretation of these measures is also the same. What differs between the GEE and standard logistic regression models are the underlying assumptions and how the parameters and their variances are estimated.

Odds ratio

$$\widehat{\text{OR}}_{(\text{DIARRHEA}=1\text{ vs. DIARRHEA}=0)} = \exp(0.2214) = 1.25$$

The odds ratio comparing symptoms of diarrhea vs. no diarrhea is calculated using the usual e to the $\hat{\beta}$ formula, yielding an estimated odds ratio of 1.25.

95% confidence interval

$$95\%\ \text{CI} = \exp[0.2214 \pm 1.96(0.8558)] = (0.23, 6.68)$$

The 95% confidence interval is calculated using the usual large-sample formula, yielding a confidence interval of (0.23, 6.68).

Wald test

$$H_0\colon \beta_3 = 0$$

$$Z = \frac{0.2214}{0.8558} = 0.259,\ P = 0.7958$$

We can test the null hypothesis that the beta coefficient for DIARRHEA is equal to zero using the Wald test, in which we divide the parameter estimate by its standard error. For the variable DIARRHEA, the Wald statistic equals 0.259. The corresponding *P*-value is 0.7958, which indicates that there is not enough evidence to reject the null hypothesis.

Standard Logistic Regression Model

```
                                          Wald
Variable   Coefficient  Std Err         p-value

INTERCEPT    −1.4362     0.6022          0.0171
BIRTHWGT     −0.0005     0.0002          0.0051
GENDER       −0.0453     0.2757          0.8694
DIARRHEA      0.7764     0.4538          0.0871
```

Responses within clusters assumed independent

Also called the "naive" model

The output for the standard logistic regression is presented for comparison. In this analysis, each observation is assumed to be independent. When there are several observations per subject, as with these data, the term "naive model" is often used to describe a model that assumes independence when responses within a cluster are likely to be correlated. For the Infant Care Study example, there are 1,203 separate outcomes across the 136 infants.

Odds ratio

$$\widehat{\text{OR}}_{(\text{DIARRHEA}=1\ \text{vs. DIARRHEA}=0)} = \exp(0.7764) = 2.17$$

Using this output, the estimated odds ratio comparing symptoms of diarrhea vs. no diarrhea is 2.17 for the naive model.

95% confidence interval

$$95\%\ \text{CI} = \exp[0.7764 \pm 1.96(0.4538)] = (0.89, 5.29)$$

The 95% confidence interval for this odds ratio is calculated to be (0.89, 5.29).

Wald test

H_0: $\beta_3 = 0$

$$Z = \frac{0.7764}{0.4538} = 1.711, \ P = 0.0871$$

The Wald test statistic for DIARRHEA in the SLR model is calculated to be 1.711. The corresponding *P*-value is 0.0871.

Comparison of analysis approaches:

1. $\widehat{\text{OR}}$ and 95% CI for DIARRHEA

	GEE model	SLR model
$\widehat{\text{OR}}$	1.25	2.17
95% CI	0.23, 6.68	0.89, 5.29

This example demonstrates that the choice of analytic approach can affect inferences made from the data. The estimates for the odds ratio and the 95% confidence interval for DIARRHEA are greatly affected by the choice of model.

2. *P*-Value of Wald test for BIRTHWGT

	GEE model	SLR model
P-Value	0.1080	0.0051

In addition, the statistical significance of the variable BIRTHWGT at the 0.05 level depends on which model is used, as the *P*-value for the Wald test of the GEE model is 0.1080, whereas the *P*-value for the Wald test of the standard logistic regression model is 0.0051.

Why these differences?

GEE model: 136 independent clusters (infants)
Naive model: 1,203 independent outcome measures

The key reason for these differences is the way the outcome is modeled. For the GEE approach, there are 136 independent clusters (infants) in the data, whereas the assumption for the standard logistic regression is that there are 1,203 independent outcome measures.

Effects of ignoring correlation structure:

- Not usually so striking
- Standard error estimates more often affected than parameter estimates
- Example shows effects on *both* standard error and parameter estimates

For many datasets, the effect of ignoring the correlation structure in the analysis is not nearly so striking. If there are differences in the resulting output from using these two approaches, it is more often the estimated standard errors of the parameter estimates rather than the parameter estimates themselves that show the greatest difference. In this example however, there are strong differences in both the parameter estimates and their standard errors.

Correlation structure:

$$\Downarrow$$

Framework for estimating:

- Correlations
- Regression coefficients
- Standard errors

To run a GEE analysis, the user specifies a correlation structure. The correlation structure provides a frame-work for the estimation of the correlation parameters, as well as estimation of the regression coefficients ($\beta_0, \beta_1, \beta_2, \ldots, \beta_p$) and their standard errors.

	Primary interest?
Regression coefficients	Yes
Correlations	Usually not

It is the regression parameters (e.g., the coefficients for DIARRHEA, BIRTWGT, and GENDER) and not the correlation parameters that typically are the parameters of primary interest.

Infant Care Study example:

AR1 autoregressive correlation structure specified

Other structures possible

Software packages that accommodate GEE analyses generally offer several choices of correlation structures that the user can easily implement. For the GEE analysis in this example, an AR1 autoregressive correlation structure was specified. Further details on the AR1 autoregressive and other correlation structures are presented later in the chapter.

In the next section (Sect. III), we present the general form of the data for a correlated analysis.

III. Data Layout

Basic data layout for correlated analysis:

K subjects

n_i responses for subject *i*

Subject	Repeat	Time	Outcome	Independent Variables			
(i)	(j)	(t_{ij})	Y_{ij}	X_{ij1}	X_{ij2}	$\cdots$	X_{ijp}
1	1	t_{11}	Y_{11}	X_{111}	X_{112}	$\cdots$	X_{11p}
1	2	t_{12}	Y_{12}	X_{121}	X_{122}	$\cdots$	X_{12p}
.	.	.	.	.	.		.
.	.	.	.	.	.		.
.	.	.	.	.	.		.
1	n_1	t_{1n_1}	Y_{1n_1}	$X_{1n_1 1}$	$X_{1n_1 2}$	$\cdots$	$X_{1n_1 p}$
.	.	.	.	.	.		.
i	1	t_{i1}	Y_{i1}	X_{i11}	X_{i12}	$\cdots$	X_{i1p}
i	2	t_{i2}	Y_{i2}	X_{i21}	X_{i22}	$\cdots$	X_{i2p}
.	.	.	.	.	.		.
.	.	.	.	.	.		.
.	.	.	.	.	.		.
i	n_i	t_{in_i}	Y_{in_i}	$X_{in_i 1}$	$X_{in_i 2}$	$\cdots$	$X_{in_i p}$
.	.	.	.	.	.		.
K	1	t_{K1}	Y_{K1}	X_{K11}	X_{K12}	$\cdots$	X_{K1p}
K	2	t_{K2}	Y_{K2}	X_{K21}	X_{K22}	$\cdots$	X_{K2p}
.	.	.	.	.	.		.
.	.	.	.	.	.		.
.	.	.	.	.	.		.
K	n_K	t_{Kn_K}	Y_{Kn_K}	$X_{Kn_K 1}$	$X_{Kn_K 2}$	$\cdots$	$X_{Kn_K p}$

The basic data layout for a correlated analysis is presented to the left. We consider a longitudinal dataset in which there are repeated measures for *K* subjects. The *i*th subject has n_i measurements recorded. The *j*th observation from the *i*th subject occurs at time t_{ij} with the outcome measured as Y_{ij}, and with *p* covariates, $X_{ij1}, X_{ij2}, \ldots, X_{ijp}$.

Subjects are not restricted to have the same number of observations (e.g., n_1 does not have to equal n_2). Also, the time interval between measurements does not have to be constant (e.g., $t_{12} - t_{11}$ does not have to equal $t_{13} - t_{12}$). Further, in a longitudinal design, a variable (t_{ij}) indicating time of measurement may be specified; however, for nonlongitudinal designs with correlated data, a time variable may not be necessary or appropriate.

The covariates (i.e., *X*s) may be time-independent or time-dependent for a given subject. For example, the race of a subject will not vary, but the daily intake of coffee could vary from day to day.

IV. Covariance and Correlation

In the sections that follow, we provide an overview of the mathematical foundation of the GEE approach. We begin by developing some of the ideas that underlie correlated analyses, including covariance and correlation.

Covariance and correlation are measures of relationships between variables.

Covariance

Population:
$\text{cov}(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$

Sample:

$$\widehat{\text{cov}}(X,Y) = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})$$

Covariance and correlation are measures that express relationships between two variables. The *covariance* of X and Y in a population is defined as the expected value, or average, of the product of X minus its mean (μ_x) and Y minus its mean (μ_y). With sample data, the covariance is estimated using the formula on the left, where $\bar{X}$ and $\bar{Y}$ are sample means in a sample of size n.

Correlation

Population: $\rho_{xy} = \dfrac{\text{cov}(X,Y)}{\sigma_x\sigma_y}$

Sample: $r_{xy} = \dfrac{\widehat{\text{cov}}(X,Y)}{s_x s_y}$

The *correlation* of X and Y in a population, often denoted by the Greek letter rho (ρ), is defined as the covariance of X and Y divided by the product of the standard deviation of X (i.e., σ_x) and the standard deviation of Y (i.e., σ_y). The corresponding sample correlation, usually denoted as r_{xy}, is calculated by dividing the sample covariance by the product of the sample standard deviations (i.e., s_x and s_y).

Correlation:

- Standardized covariance
- Scale free

X_1 = height (in feet)

X_2 = height (in inches)

Y = weight

$\text{cov}(X_2, Y) = 12\,\text{cov}(X_1, Y)$

BUT

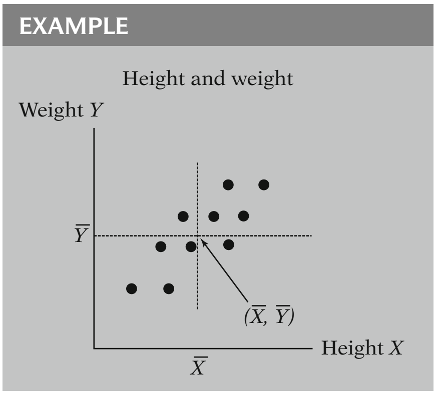
$\rho_{x_2y} = \rho_{x_1y}$

The correlation is a standardized measure of covariance in which the units of X and Y are the standard deviations of X and Y, respectively. The actual units used for the value of variables affect measures of covariance but not measures of correlation, which are scale-free. For example, the covariance between height and weight will increase by a factor of 12 if the measure of height is converted from feet to inches, but the correlation between height and weight will remain unchanged.

Positive correlation

On average, as X gets larger, Y gets larger; or, as X gets smaller, Y gets smaller.

EXAMPLE

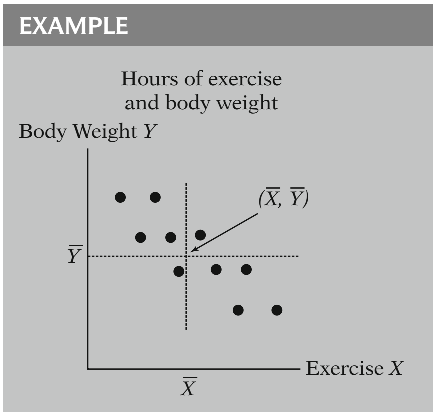


A positive correlation between X and Y means that larger values of X, on average, correspond with larger values of Y, whereas smaller values of X correspond with smaller values of Y. For example, persons who are above mean height will be, on average, above mean weight, and persons who are below mean height will be, on average, below mean weight. This implies that the correlation between individuals' height and weight measurements is positive. This is not to say that there cannot be tall people of below average weight or short people of above average weight. Correlation is a measure of average, even though there may be variation among individual observations. Without any additional knowledge, we would expect a person 6 ft tall to weigh more than a person 5 ft tall.

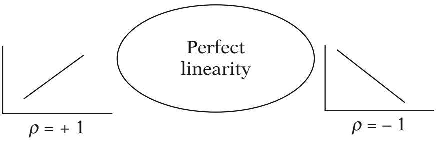
Negative correlation

On average, as X gets larger, Y gets smaller; or as X gets smaller, Y gets larger.

EXAMPLE



A negative correlation between X and Y means that larger values of X, on average, correspond with smaller values of Y, whereas smaller values of X correspond with larger values of Y. An example of negative correlation might be between hours of exercise per week and body weight. We would expect, on average, people who exercise more to weigh less, and conversely, people who exercise less to weigh more. Implicit in this statement is the control of other variables such as height, age, gender, and ethnicity.



The possible values of the correlation of X and Y range from negative 1 to positive 1. A correlation of negative 1 implies that there is a perfect negative linear relationship between X and Y, whereas a correlation of positive 1 implies a perfect positive linear relationship between X and Y.

Perfect linear relationship

$Y = \beta_0 + \beta_1 X$, for a given X

X and Y independent $\Rightarrow \rho = 0$

BUT

$$\rho = 0 \Rightarrow \begin{cases} X \text{ and } Y \text{ independent} \\ \textbf{or} \\ X \text{ and } Y \text{ have nonlinear} \\ \text{relationship} \end{cases}$$

By a perfect linear relationship we mean that, given a value of X, the value of Y can be exactly ascertained from that linear relationship of X and Y (i.e., $Y = \beta_0 + \beta_1 X$ where β_0 is the intercept and β_1 is the slope of the line). If X and Y are independent, then their correlation will be zero. The reverse does not necessarily hold. A zero correlation may also result from a nonlinear association between X and Y.

Correlations on same variable

$(Y_1, Y_2, \ldots, Y_n)$

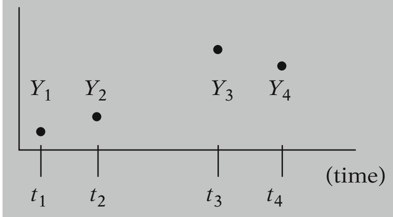
$\rho_{Y_1Y_2}, \rho_{Y_1Y_3}, \ldots,$ etc.

We have been discussing correlation in terms of two different variables such as height and weight. We can also consider correlations between repeated observations (Y_1, Y_2, $\ldots$, Y_n) on the same variable Y.

EXAMPLE

Systolic blood pressure on same individual over time

Expect $\rho_{Y_jY_k} > 0$ for some j, k
Also,



Expect $\rho_{Y_1Y_2}$ or $\rho_{Y_3Y_4} >$
$\rho_{Y_1Y_3}$, $\rho_{Y_1Y_4}$, $\rho_{Y_2Y_3}$, $\rho_{Y_2Y_4}$

Consider a study in which each subject has several systolic blood pressure measurements over a period of time. We might expect a positive correlation between pairs of blood pressure measurements from the same individual (Y_j, Y_k).

The correlation might also depend on the time period between measurements. Measurements 5 min apart on the same individual might be more highly correlated than measurements 2 years apart.

Correlations between dichotomous variables may also be considered.

This discussion can easily be extended from continuous variables to dichotomous variables. Suppose a study is conducted examining daily inhaler use by patients with asthma. The dichotomous outcome is coded 1 for the event (use) and 0 for no event (no use). We might expect a positive correlation between pairs of responses from the same subject (Y_j, Y_k).

EXAMPLE

Daily inhaler use (1 = yes, 0 = no) on same individual over time

Expect $\rho_{Y_jY_k} > 0$ for same subject

V. Generalized Linear Models

General form of many statistical models:

$Y = f(X_1, X_2, \ldots, X_p) + \epsilon,$

where: Y is random
$X_1, X_2, \ldots, X_p$ are fixed
ϵ is random

For many statistical models, including logistic regression, the predictor variables (i.e., independent variables) are considered fixed and the outcome, or response (i.e., dependent variable), is considered random. A general formulation of this idea can be expressed as $Y = f(X_1, X_2, \ldots, X_p) + \epsilon$ where Y is the response variable, $X_1, X_2, \ldots, X_p$ are the predictor variables, and ϵ represents random error. In this framework, the model for Y consists of a fixed component $[f(X_1, X_2, \ldots, X_p)]$ and a random component (ϵ).

Specify:

1. A function (f) for the fixed predictors, e.g., linear
2. A distribution for the random error (ϵ), e.g., N(0,1)

A function (f) for the fixed predictors and a distribution for the random error (ϵ) are specified.

GLM models include:

Logistic regression
Linear regression
Poisson regression

GEE models are extensions of GLM

Logistic regression belongs to a class of models called generalized linear models (GLM). Other models that belong to the class of GLM include linear and Poisson regression. For correlated analyses, the GLM framework can be extended to a class of models called generalized estimating equations (GEE) models. Before discussing correlated analyses using GEE, we shall describe GLM.

GLM: a generalization of the classical linear model

Linear regression
Outcome:
- Continuous
- Normal distribution

GLM are a natural generalization of the classical linear model (McCullagh and Nelder, 1989). In classical linear regression, the outcome is a continuous variable, which is often assumed to follow a normal distribution. The mean response is modeled as linear with respect to the regression parameters.

Logistic regression
Outcome:
- Dichotomous
- Binomial distribution: $E(Y) = \mu = \mathrm{P}(Y = 1)$

In standard logistic regression, the outcome is a dichotomous variable. Dichotomous outcomes are often assumed to follow a binomial distribution, with an expected value (or mean, μ) equal to a probability [e.g., $\mathrm{P}(Y = 1)$].

Logistic regression used to model

$$\mathrm{P}(Y = 1 \mid X_1, X_2, \ldots, X_p)$$

It is this probability that is modeled in logistic regression.

Exponential family distributions include:
Binomial
Normal
Poisson
Exponential
Gamma

The binomial distribution belongs to a larger class of distributions called the exponential family. Other distributions belonging to the exponential family are the normal, Poisson, exponential, and gamma distributions. These distributions can be written in a similar form and share important properties.

Generalized linear model

$$g(\mu) = \beta_0 + \sum_{h=1}^{p} \beta_h X_h,$$

where: μ is the mean response $E(Y)$
$g(\mu)$ is a function of the mean

Let μ represent the mean response $E(Y)$, and $g(\mu)$ represent a function of the mean response. A generalized linear model with p independent variables can be expressed as $g(\mu)$ equals β_0 plus the summation of the p independent variables times their beta coefficients.

Three components for GLM:

1. Random component
2. Systematic component
3. Link function

There are three components that comprise GLM: (1) a random component, (2) a systematic component, and (3) the link function. These components are described as follows:

1. **Random component**

   Y follows a distribution from the exponential family

1. The *random component* requires the outcome (Y) to follow a distribution from the exponential family. This criterion is met for a logistic regression (unconditional) since the response variable follows a binomial distribution, which is a member of the exponential family.

2. **Systematic component**

   The Xs are combined in the model linearly, (i.e., $\beta_0 + \Sigma\beta_h X_h$)

   Logistic model:

$$P(\mathbf{X}) = \frac{1}{1 + \exp[-(\underset{\substack{\nwarrow \\ \text{linear component}}}{\beta_0} + \Sigma\, \beta_h X_h]}$$

2. The *systematic component* requires that the Xs be combined in the model as a linear function ($\beta_0 + \Sigma\beta_h X_h$) of the parameters. This portion of the model is not random. This criterion is met for a logistic model, since the model form contains a linear component in its denominator.

3. **Link function:**

$g(\mu) = \beta_0 + \sum \beta_h X_h$

g ''links'' $E(Y)$ with $\beta_0 + \sum \beta_h X_h$

3. The *link function* refers to that function of the mean response, $g(\mu)$, that is modeled linearly with respect to the regression parameters. This function serves to "link" the mean of the random response and the fixed linear set of parameters.

Logistic regression (logit link)

$$g(\mu) = \log\left[\frac{\mu}{1-\mu}\right] = \text{logit}(\mu)$$

For logistic regression, the *log odds* (or *logit*) of the outcome is modeled as linear in the regression parameters. Thus, the link function for logistic regression is the logit function [i.e., $g(\mu)$ equals the log of the quantity μ divided by 1 minus μ].

Alternate formulation

Inverse of link function $= g^{-1}$ satisfies

$$g^{-1}(g(\mu)) = \mu$$

Inverse of logit function in terms of $(\mathbf{X}, \beta)$

$$g^{-1}(\mathbf{X}, \boldsymbol{\beta}) = \mu = \frac{1}{1 + \exp\left[-\left(\alpha + \sum_{h=1}^{p} \beta_h \mathrm{X}_h\right)\right]},$$

where

$$g(\mu) = \text{logit}\,\mathrm{P}(D = 1 \mid \mathbf{X}) = \beta_0 + \sum_{h=1}^{p} \beta_h X_h$$

Alternately, one can express GLM in terms of the *inverse* of the link function (g^{-1}), which is the mean μ. In other words, $g^{-1}(g(\mu)) = \mu$. This inverse function is modeled in terms of the predictors ($\mathbf{X}$) and their coefficients (β) (i.e., $g^{-1}(\mathbf{X}, \beta)$). For logistic regression, the inverse of the logit link function is the familiar logistic model of the probability of an event, as shown on the left. Notice that this modeling of the mean (i.e., the inverse of the link function) is not a linear model. It is the *function* of the mean (i.e., the link function) that is modeled as linear in GLM.

GLM:

- Uses ML estimation
- Requires likelihood function L where

  $$L = \prod_{i=1}^{K} L_i$$

  (assumes Y_i are independent)

GLM uses maximum likelihood methods to estimate model parameters. This requires knowledge of the likelihood function (L), which, in turn, requires that the distribution of the response variable be specified.

If the responses are independent, the likelihood can be expressed as the product of each observation's contribution (L_i) to the likelihood.

If Y_i not independent and not normal

$\Downarrow$

L complicated or intractable

However, if the responses are not independent, then the likelihood can become complicated, or intractable.

If Y_i not independent but MV normal

$$\Downarrow$$

L specified

If Y_i not independent and *not* MV normal

$$\Downarrow$$

Quasi-likelihood theory

For nonindependent outcomes whose joint distribution is multivariate (MV) normal, the likelihood is relatively straightforward, since the multivariate normal distribution is completely specified by the means, variances, and all of the pairwise covariances of the random outcomes. This is typically *not* the case for other multivariate distributions in which the outcomes are *not* independent. For these circumstances, quasi-likelihood theory offers an alternative approach for model development.

Quasi-likelihood:

- No likelihood
- Specify mean variance relationship
- Foundation of GEE

Quasi-likelihood methods have many of the same desirable statistical properties that maximum likelihood methods have, but the full likelihood does not need to be specified. Rather, the relationship between the mean and variance of each response is specified. Just as the maximum likelihood theory lays the foundation for GLM, the quasi-likelihood theory lays the foundation for GEE models.

VI. GEE Models

GEE: class of models for correlated data Link function ***g*** modeled as

$$g(\mu) = \beta_0 + \sum_{h=1}^{p} \beta_h X_h$$

GEE represent a class of models that are often utilized for data in which the responses are correlated (Liang and Zeger, 1986). GEE models can be used to account for the correlation of continuous or categorical outcomes. As in GLM, a function of the mean $g(\mu)$, called the *link function*, is modeled as linear in the regression parameters.

For $Y(0,1) \Rightarrow$ logit link

$$g(\mu) = \text{logit } \text{P}(Y = 1 \,|\, \mathbf{X}) = \beta_0 + \sum_{h=1}^{p} \beta_h X_h$$

For a dichotomous outcome, the logit link is commonly used. For this case, $g(\mu)$ equals logit (P), where P is the probability that $Y = 1$. If there are p independent variables, this can be expressed as: logit $\text{P}(Y = 1 \,|\, \mathbf{X})$ equals β_0 plus the summation of the p independent variables times their β coefficients.

Correlated vs. independent

- Identical model

  but
- Different assumptions

The logistic model for correlated data looks identical to the standard logistic model. The difference is in the underlying assumptions of the model, including the presence of correlation, and the way in which the parameters are estimated.

GEE:

- Generalization of quasi-likelihood
- Specify a “working” correlation structure for within-cluster correlations
- Assume independence between clusters

GEE is a generalization of quasi-likelihood estimation, so the joint distribution of the data need not be specified. For clustered data, the user specifies a “working” correlation structure for describing how the responses within clusters are related to each other. Between clusters, there is an assumption of independence.

EXAMPLE

Asthma patients followed 7 days

Y: daily inhaler use (0,1)
E: pollen level
Cluster: asthma patient

Y_i *within* subjects correlated

but

Y_i *between* subjects independent

For example, suppose 20 asthma patients are followed for a week and keep a daily diary of inhaler use. The response (Y) is given a value of 1 if a patient uses an inhaler on a given day and 0 if there is no use of an inhaler on that day. The exposure of interest is daily pollen level. In this analysis, each subject is a cluster. It is reasonable to expect that outcomes (i.e., daily inhaler use) are positively correlated within observations from the same subject but independent between different subjects.

VII. Correlation Structure

Correlation and covariance summarized as square matrices

The correlation and the covariance between measures are often summarized in the form of a square matrix (i.e., a matrix with equal numbers of rows and columns). We use simple matrices in the following discussion; however, a background in matrix operations is not required for an understanding of the material.

Covariance matrix for Y_1 and Y_2

$$\mathbf{V} = \begin{bmatrix} \operatorname{var}(Y_1) & \operatorname{cov}(Y_1, Y_2) \\ \operatorname{cov}(Y_1, Y_2) & \operatorname{var}(Y_2) \end{bmatrix}$$

For simplicity consider two observations, Y_1 and Y_2. The covariance matrix for just these two observations is a 2 × 2 matrix ($\mathbf{V}$) of the form shown at left. We use the conventional matrix notation of bold capital letters to identify individual matrices.

Corresponding 2 × 2 correlation matrix

$$\mathbf{C} = \begin{bmatrix} 1 & \text{corr}(Y_1, Y_2) \\ \text{corr}(Y_1, Y_2) & 1 \end{bmatrix}$$

The corresponding 2 × 2 correlation matrix (**C**) is also shown at left. Note that the covariance between a variable and itself is the variance of that variable [e.g., $\text{cov}(Y_1, Y_1) = \text{var}(Y_1)$], so that the correlation between a variable and itself is 1.

Diagonal matrix: has 0 in all nondiagonal entries.

A *diagonal matrix* has a 0 in all nondiagonal entries.

Diagonal 2 × 2 matrix with variances on diagonal

$$\mathbf{D} = \begin{bmatrix} \text{var}(Y_1) & 0 \\ 0 & \text{var}(Y_2) \end{bmatrix}$$

A 2 × 2 diagonal matrix (**D**) with the variances along the diagonal is of the form shown at left.

Can extend to $N \times N$ matrices

Matrices symmetric: $(i,\ j) = (j,\ i)$ element

$$\text{cov}(Y_1, Y_2) = \text{cov}(Y_2, Y_1)$$
$$\text{corr}(Y_1, Y_2) = \text{corr}(Y_2, Y_1)$$

The definitions of **V**, **C**, and **D** can be extended from 2 × 2 matrices to $N \times N$ matrices. A *symmetric matrix* is a square matrix in which the (i, j) element of the matrix is the same value as the (j, i) element. The covariance of (Y_i, Y_j) is the same as the covariance of (Y_j, Y_i); thus the covariance and correlation matrices are symmetric matrices.

Relationship between covariance and correlation expressed as

$$\text{cov}(Y_1, Y_2) = \sqrt{\text{var}(Y_1)}[\text{corr}(Y_1, Y_2)]\sqrt{\text{var}(Y_2)}$$

The covariance between Y_1 and Y_2 equals the standard deviation of Y_1, times the correlation between Y_1 and Y_2, times the standard deviation of Y_2.

Matrix version: $\mathbf{V} = \mathbf{D}^{\frac{1}{2}}\mathbf{C}\mathbf{D}^{\frac{1}{2}}$,

where $\mathbf{D}^{\frac{1}{2}} \times \mathbf{D}^{\frac{1}{2}} = \mathbf{D}$

The relationship between covariance and correlation can be similarly expressed in terms of the matrices **V**, **C**, and **D** as shown on the left.

Logistic regression

$$\mathbf{D} = \begin{bmatrix} \mu_1(1-\mu_1) & 0 \\ 0 & \mu_2(1-\mu_2) \end{bmatrix},$$

where

$$\text{var}(Y_i) = \mu_i(1-\mu_i)$$
$$\mu_i = g^{-1}(\mathbf{X}, \beta)$$

For logistic regression, the variance of the response Y_i equals μ_i times $(1 - \mu_i)$. The corresponding diagonal matrix (**D**) has $\mu_i(1 - \mu_i)$ for the diagonal elements and 0 for the off-diagonal elements. As noted earlier, the mean (μ_i) is expressed as a function of the covariates and the regression parameters $[g^{-1}(\mathbf{X}, \beta)]$.

EXAMPLE

Three subjects; four observations each
Within-cluster correlation between jth and kth response from subject $i = \rho_{ijk}$
Between-subject correlations $= 0$

$$\begin{bmatrix}
1 & \rho_{112} & \rho_{113} & \rho_{114} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\rho_{112} & 1 & \rho_{123} & \rho_{124} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\rho_{113} & \rho_{123} & 1 & \rho_{134} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\rho_{114} & \rho_{124} & \rho_{134} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & \rho_{212} & \rho_{213} & \rho_{214} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \rho_{212} & 1 & \rho_{223} & \rho_{224} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \rho_{213} & \rho_{223} & 1 & \rho_{234} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \rho_{214} & \rho_{224} & \rho_{234} & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \rho_{312} & \rho_{313} & \rho_{314} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho_{312} & 1 & \rho_{323} & \rho_{324} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho_{313} & \rho_{323} & 1 & \rho_{334} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho_{314} & \rho_{324} & \rho_{334} & 1
\end{bmatrix}$$

blocks

We illustrate the form of the correlation matrix in which responses are correlated within subjects and independent between subjects. For simplicity, consider a dataset with information on only three subjects in which there are four responses recorded for each subject. There are 12 observations (3 times 4) in all. The correlation between responses from two different subjects is 0, whereas the correlation between responses from the same subject (i.e., the jth and kth response from subject i) is ρ_{ijk}.

Block diagonal matrix: subject-specific correlation matrices form blocks (B_i)

$$\begin{bmatrix} \mathbf{B_1} & & \mathbf{0} \\ & \mathbf{B_2} & \\ \mathbf{0} & & \mathbf{B_3} \end{bmatrix} \text{ where } B_i = i\text{th block}$$

This correlation matrix is called a *block diagonal matrix*, where subject-specific correlation matrices are the blocks along the diagonal of the matrix.

EXAMPLE

18 ρs (6 per cluster/subject) but 12 observations

$$\downarrow$$

Subject i: $\{\rho_{i12}, \rho_{i13}, \rho_{i14}, \rho_{i23}, \rho_{i24}, \rho_{i34}\}$

The correlation matrix in the preceding example contains 18 correlation parameters (6 per cluster) based on only 12 observations. In this setting, each subject has his or her own distinct set of correlation parameters.

# parameters > # observations
$\Rightarrow \hat{\beta}_i$ not valid

GEE approach: common set of ρs for each subject:

Subject i: $\{\rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34}\}$

If there are more parameters to estimate than observations in the dataset, then the model is overparameterized and there is not enough information to yield valid parameter estimates. To avoid this problem, *the GEE approach requires that each subject have a common set of correlation parameters*. This reduces the number of correlation parameters substantially. This type of correlation matrix is presented at the left.

EXAMPLE

3 subjects; 4 observations each

$$\begin{bmatrix}
1 & \rho_{12} & \rho_{13} & \rho_{14} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\rho_{12} & 1 & \rho_{23} & \rho_{24} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\rho_{13} & \rho_{23} & 1 & \rho_{34} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\rho_{14} & \rho_{24} & \rho_{34} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & \rho_{12} & \rho_{13} & \rho_{14} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \rho_{12} & 1 & \rho_{23} & \rho_{24} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \rho_{13} & \rho_{23} & 1 & \rho_{34} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \rho_{14} & \rho_{24} & \rho_{34} & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \rho_{12} & \rho_{13} & \rho_{14} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho_{12} & 1 & \rho_{23} & \rho_{24} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho_{13} & \rho_{23} & 1 & \rho_{34} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho_{14} & \rho_{24} & \rho_{34} & 1
\end{bmatrix}$$

Now only 6 ρs for 12 observations:
$\downarrow$ # ρs by factor of 3 (= # subjects)

There are now 6 correlation parameters (ρ_{jk}) for 12 observations of data. Giving each subject a common set of correlation parameters reduced the number by a factor of 3 (18 to 6).

In general, for K subjects:
$\rho_{ijk} \Rightarrow \rho_{jk}$: # of ρs $\downarrow$ by factor of K

In general, a common set of correlation parameters for K subjects reduces the number of correlation parameters by a factor of K.

Example above: *unstructured* correlation structure

Next section shows other structures.

The correlation structure presented above is called *unstructured*. Other correlation structures, with stronger underlying assumptions, reduce the number of correlation parameters even further. Various types of correlation structure are presented in the next section.

VIII. Different Types of Correlation Structure

Examples of correlation structures:

- Independent
- Exchangeable
- AR1 autoregressive
- Stationary m-dependent
- Unstructured
- Fixed

We present a variety of correlation structures that are commonly considered when performing a correlated analysis. These correlation structures are as follows: independent, exchangeable, AR1 autoregressive, stationary *m*-dependent, unstructured, and fixed. Software packages that accommodate correlated analyses typically allow the user to specify the correlation structure before providing estimates of the correlation parameters.

Independent

Assumption: responses uncorrelated within clusters

Matrix for a given cluster is the *identity matrix*.

With five responses per cluster

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Independent correlation structure.

The assumption behind the use of the independent correlation structure is that responses are uncorrelated within a cluster. The correlation matrix for a given cluster is just the *identity matrix*. The identity matrix has a value of 1 along the main diagonal and a 0 off the diagonal. The correlation matrix to the left is for a cluster that has five responses.

Exchangeable

Assumption: any two responses within a cluster have same correlation (ρ)

With five responses per cluster

$$\begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix}$$

Exchangeable correlation structure.

The assumption behind the use of the exchangeable correlation structure is that any two responses within a cluster have the same correlation (ρ). The correlation matrix for a given cluster has a value of 1 along the main diagonal and a value of ρ off the diagonal. The correlation matrix to the left is for a cluster that has five responses.

Only one ρ estimated

As in all correlation structures used for GEE analyses, the same set of correlation parameters are assumed for modeling each cluster. For the exchangeable correlation structure, this means that there is only one correlation parameter to be estimated.

Order of observations within a cluster is arbitrary.

Can exchange positions of observations.

$K = 14$ schools

$n_i = \#$ students from school i

$$\sum_{i=1}^{K} n_i = 237$$

School i: exchange order $\# 2 \leftrightarrow \# 9$

$$\Downarrow$$

Will not affect analysis

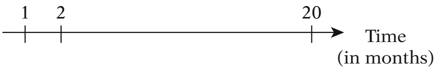
A feature of the exchangeable correlation structure is that the order of observations within a cluster is arbitrary. For example, consider a study in which there is a response from each of 237 students representing 14 different high schools. It may be reasonable to assume that responses from students who go to the same school are correlated. However, for a given school, we would not expect the correlation between the response of student #1 and student #2 to be different from the correlation between the response of student #1 and student #9. We could therefore *exchange* the order (the position) of student #2 and student #9 and not affect the analysis.

Number of responses (n_i) can vary by i

It is not required that there be the same number of responses in each cluster. We may have 10 students from one school and 15 students from a different school.

Autoregressive

Assumption: correlation depends on interval of time between responses



$\rho_{1,2} > \rho_{1,20}$

Autoregressive correlation structure: An autoregressive correlation structure is generally applicable for analyses in which there are repeated responses *over time* within a given cluster. The assumption behind an autoregressive correlation structure is that the correlation between responses depends on the interval of time between responses. For example, the correlation is assumed to be greater for responses that occur 1 month apart rather than 20 months apart.

AR1

Special case of autoregressive

Assumption: Y at t_1 and t_2:

$$\rho_{t_1,t_2} = \rho^{|t_1 - t_2|}$$

AR1 is a special case of an autoregressive correlation structure. AR1 is widely used because it assumes only one correlation parameter and because software packages readily accommodate it. The AR1 assumption is that the correlation between any two responses from the same subject equals a baseline correlation (ρ) raised to a power equal to the absolute difference between the times of the responses.

Cluster with four responses at time $t = 1, 2, 3, 4$

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

The correlation matrix to the left is for a cluster that has four responses taken at time $t = 1, 2, 3, 4$.

Cluster with four responses at time $t = 1, 6, 7, 10$

$$\begin{bmatrix} 1 & \rho^5 & \rho^6 & \rho^9 \\ \rho^5 & 1 & \rho^5 & \rho^6 \\ \rho^6 & \rho^5 & 1 & \rho^5 \\ \rho^9 & \rho^6 & \rho^5 & 1 \end{bmatrix}$$

Contrast this to another example of an AR1 correlation structure for a cluster that has four responses taken at time $t = 1, 6, 7, 10$. In each example, the power to which rho (ρ) is raised is the difference between the times of the two responses.

With AR1 structure, only one ρ
BUT
Order within cluster *not* arbitrary

As with the exchangeable correlation structure, the AR1 structure has just one correlation parameter. In contrast to the exchangeable assumption, the order of responses within a cluster is not arbitrary, as the time interval is also taken into account.

Stationary m-dependent

Assumption:
Correlations k occasions apart same for $k = 1, 2, \ldots, m$
Correlations $> m$ occasions apart $= 0$

Stationary 2-dependent, cluster with six responses ($m = 2, n_i = 6$)

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & 0 & 0 & 0 \\ \rho_1 & 1 & \rho_1 & \rho_2 & 0 & 0 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & 0 \\ 0 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ 0 & 0 & \rho_2 & \rho_1 & 1 & \rho_1 \\ 0 & 0 & 0 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

Stationary m-dependent structure $\Rightarrow m$ distinct ρs

Unstructured

Cluster with four responses
$\# \rho = 4(3)/2 = 6$

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}$$

$$n \text{ responses} \Downarrow n(n-1)/2 \text{ distinct } \rho\text{s},$$
i.e. $\rho_{jk} \neq \rho_{j'k'}$ unless $j = j'$ and $k = k'$

$\rho_{12} \neq \rho_{34}$ even if $t_2 - t_1 = t_4 - t_3$

Stationary m-dependent correlation structure:

The assumption behind the use of the stationary m-dependent correlation structure is that correlations k occasions apart are the same for $k = 1, 2, \ldots, m$, whereas correlations more than m occasions apart are zero.

The correlation matrix to the left illustrates a stationary 2-dependent correlation structure for a cluster that has six responses. A stationary 2-dependent correlation structure has two correlation parameters.

In general, a stationary m-dependent correlation structure has m distinct correlation parameters. The assumption here is that responses within a cluster are uncorrelated if they are more than m units apart.

Unstructured correlation structure:

In an unstructured correlation structure there are less constraints on the correlation parameters. The correlation matrix to the left is for a cluster that has four responses and six correlation parameters.

In general, for a cluster that has n responses, there are $n(n-1)/2$ correlation parameters. If there are a large number of correlation parameters to estimate, the model may be unstable and results unreliable.

An unstructured correlation structure has a separate correlation parameter for each pair of observations (j, k) within a cluster, even if the time intervals between the responses are the same. For example, the correlation between the first and second responses of a cluster is not assumed to be equal to the correlation between the third and fourth responses.

$\rho_{ijk} = \rho_{i'jk}$ if $i \neq i'$

$$\rho_{\text{A}12} = \rho_{\text{B}12} = \rho_{12}$$

different clusters

Order $\{Y_{i1}, Y_{i2}, \ldots, Y_{ik}\}$ *not* arbitrary (e.g., cannot switch $Y_{\text{A}1}$ and $Y_{\text{A}4}$ unless all Y_{i1} and Y_{i4} switched).

Like the other correlation structures, the same set of correlation parameters are used for each cluster. Thus, the correlation between the first and second responses for cluster A is the same as the correlation between the first and second response for cluster B. This means that the order of responses for a given cluster is not arbitrary for an unstructured correlation structure. If we exchange the first and fourth responses of cluster i, it does affect the analysis, unless we also exchange the first and fourth responses for all the clusters.

Fixed

Fixed correlation structure.

User specifies fixed values for ρ.

$\rho = 0.1$ for first and fourth responses; 0.3 otherwise

$$\begin{bmatrix} 1.0 & 0.3 & 0.3 & 0.1 \\ 0.3 & 1.0 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1.0 & 0.3 \\ 0.1 & 0.3 & 0.3 & 1.0 \end{bmatrix}$$

Some software packages allow the user to select fixed values for the correlation parameters. Consider the correlation matrix presented on the left. The correlation between the first and fourth responses of each cluster is fixed at 0.1; otherwise, the correlation is fixed at 0.3.

No ρ estimated.

For an analysis that uses a fixed correlation structure, there are no correlation parameters to estimate since the values of the parameters are chosen before the analysis is performed.

Choice of structure not always clear.

Selection of a "working" correlation structure is at the discretion of the researcher. Which structure best describes the relationship between correlations is not always clear from the available evidence. For large samples, the estimates of the standard errors of the parameters are more affected by the choice of correlation structure than the estimates of the parameters themselves.

IX. Empirical and Model-Based Variance Estimators

In the next section, we describe two variance estimators that can be obtained for the fitted regression coefficients – empirical and model-based estimators. In addition, we discuss the effect of misspecification of the correlation structure on those estimators.

GEE estimates have desirable $\underbrace{\textit{asymptotic}}$ properties.

$K \to \infty$ (i.e., K "large"), where $K = \#$ clusters

"Large" is subjective

Maximum likelihood estimates in GLM are appealing because they have desirable asymptotic statistical properties. Parameter estimates derived from GEE share some of these properties. By asymptotic, we mean "as the number of clusters approaches infinity". This is a theoretical concept since the datasets that we are considering have a finite sample size. Rather, we can think of these properties as holding for large samples. Nevertheless, the determination of what constitutes a "large" sample is somewhat subjective.

Two statistical properties of GEE estimates (if model correct):

1. **Consistent**

   $\hat{\beta} \to \beta \quad \text{as } K \to \infty$

2. **Asymptotically normal**

   $\hat{\beta} \sim \text{normal as } K \to \infty$

If a GEE model is correctly specified, then the resultant regression parameter estimates have two important statistical properties: (1) the estimates are *consistent* and (2) the distribution of the estimates is asymptotically normal. A consistent estimator is a parameter estimate that approaches the true parameter value in probability. In other words, as the number of clusters becomes sufficiently large, the difference between the parameter estimate and the true parameter approaches zero. Consistency is an important statistical property since it implies that the method will asymptotically arrive at the correct answer. The asymptotic normal property is also important since knowledge of the distribution of the parameter estimates allows us to construct confidence intervals and perform statistical tests.

Asymptotic normal property allows:
- Confidence intervals
- Statistical tests

To correctly specify a GEE model:

- Specify correct $g(\mu)$
- Specify correct $\mathbf{C}_i$

To correctly specify a GLM or GEE model, one must correctly model the mean response [i.e., specify the correct link function $g(\mu)$ and use the correct covariates]. Otherwise, the parameter estimates will not be consistent. An additional issue for GEE models is whether the correlation structure is correctly specified by the working correlation structure ($\mathbf{C}_i$).

$\hat{\beta}_h$ consistent even if $\mathbf{C}_i$ misspecified

but

$\hat{\beta}_h$ more efficient if $\mathbf{C}_i$ correct

A key property of GEE models is that parameter estimates for the regression coefficients are consistent even if the correlation structure is misspecified. However, it is still preferable for the correlation structure to be correctly specified. There is less propensity for error in the parameter estimates (i.e., smaller variance) if the correlation structure is correctly specified. Estimators are said to be more *efficient* if the variance is smaller.

To construct CIs, need $\widehat{\text{var}}(\hat{\beta})$

Two types of variance estimators:

- Model-based
- Empirical

No effect on $\hat{\beta}$

Effect on $\widehat{\text{var}}(\hat{\beta})$

For the construction of confidence intervals (CIs), it is not enough to know that the parameter estimates are asymptotically normal. In addition, we need to estimate the variance of the parameter estimates (not to be confused with the variance of the outcome). For GEE models, there are two types of variance estimator, called *model-based* and *empirical*, that can be obtained for the fitted regression coefficients. The choice of which estimator is used has no effect on the parameter estimate ($\hat{\beta}$), but rather the effect is on the estimate of its variance [$\widehat{\text{var}}(\hat{\beta})$].

Model-based variance estimators:

- Similar in form to variance estimators in GLM
- Consistent only if $\mathbf{C_i}$ correctly specified

Model-based variance estimators are of a similar form as the variance estimators in a GLM, which are based on maximum likelihood theory. Although the likelihood is never formulated for GEE models, model-based variance estimators are consistent estimators, but only if the correlation structure is correctly specified.

Empirical (robust) variance estimators:

- An adjustment of model-based estimators
- Uses observed ρ_{jk} between responses
- *Consistent even if* $\boldsymbol{C}_i$ *misspecified*

↑

Advantage of empirical estimator

Empirical (robust) variance estimators are an adjustment of model-based estimators (see Liang and Zeger, 1986). Both the model-based approach and the empirical approach make use of the working correlation matrix. However, the empirical approach also makes use of the observed correlations between responses in the data. The advantage of using the empirical variance estimator is that it *provides a consistent estimate of the variance even if the working correlation is not correctly specified*.

Estimation of β vs. estimation of var($\hat{\beta}$)

- β is estimated by $\hat{\beta}$
- $\text{var}(\hat{\beta})$ is estimated by $\widehat{\text{var}}(\hat{\beta})$

The true value of β *does not depend* on the study

The true value of $\text{var}(\hat{\beta})$ *does depend* on the study design and the type of analysis

There is a conceptual difference between the estimation of a regression coefficient and the estimation of its variance [$\widehat{\text{var}}(\hat{\beta})$]. The regression coefficient, β, is assumed to exist whether a study is implemented or not. The distribution of $\hat{\beta}$, on the other hand, depends on characteristics of the study design and the type of analysis performed. For a GEE analysis, the distribution of $\hat{\beta}$ depends on such factors as the true value of β, the number of clusters, the number of responses within the clusters, the true correlations between responses, and the working correlation structure specified by the user. Therefore, the *true* variance of $\hat{\beta}$ (and not just its estimate) depends, in part, on the choice of a working correlation structure.

Choice of working correlation structure

⇒ affects **true** variance of $\hat{\beta}$

Empirical estimator generally recommended.

Reason: robust to misspecification of correlation structure

Preferable to specify working correlation structure close to actual one:

- More efficient estimate of β
- More reliable estimate of var$(\hat{\beta})$ if number of clusters is small

For the estimation of the variance of $\hat{\beta}$ in the GEE model, the empirical estimator is generally recommended over the model-based estimator since it is more robust to misspecification of the correlation structure. This may seem to imply that if the empirical estimator is used, it does not matter which correlation structure is specified. However, choosing a working correlation that is closer to the actual one is preferable since there is a gain in efficiency. Additionally, since consistency is an asymptotic property, if the number of clusters is small, then even the empirical variance estimate may be unreliable (e.g., may yield incorrect confidence intervals) if the correlation structure is misspecified.

X. Statistical Tests

In SLR, three tests of significance of $\hat{\beta}_h$s:

- Likelihood ratio test
- Score test
- Wald test

The *likelihood ratio test*, the *Wald test*, and the *Score test* can each be used to test the statistical significance of regression parameters in a standard logistic regression (SLR). The formulation of the likelihood ratio statistic relies on the likelihood function. The formulation of the Score statistic relies on the score function, (i.e., the partial derivatives of the log likelihood). (Score functions are described in Sect. XI.) The formulation of the Wald test statistic relies on the parameter estimate and its variance estimate.

In GEE models, two tests of $\hat{\beta}_h$:

- Score test
- Wald test

~~likelihood ratio test~~

For GEE models, the likelihood ratio test cannot be used since a likelihood is never formulated. However, there is a generalization of the Score test designed for GEE models. The test statistic for this Score test is based on the generalized estimating "score-like" equations that are solved to produce parameter estimates for the GEE model. (These "score-like" equations are described in Sect. XI.) The Wald test can also be used for GEE models since parameter estimates for GEE models are asymptotically normal.

To test several $\hat{\beta}_h$ simultaneously use

- Score test
- Generalized Wald test

The Score test, as with the likelihood ratio test, can be used to test several parameter estimates simultaneously (i.e., used as a chunk test). There is also a generalized Wald test that can be used to test several parameter estimates simultaneously.

Under H_0, test statistics approximate χ^2 with df = number of parameters tested.

The test statistics for both the Score test and the generalized Wald test are similar to the likelihood ratio test in that they follow an approximate chi-square distribution under the null with the degrees of freedom equal to the number of parameters that are tested. When testing a single parameter, the generalized Wald test statistic reduces to the familiar form $\hat{\beta}_h$ divided by the estimated standard error of $\hat{\beta}_h$.

To test one $\hat{\beta}_h$, the Wald test statistic is of the familiar form

$$Z = \frac{\hat{\beta}_h}{s_{\hat{\beta}_h}}$$

The use of the Score test, Wald test, and generalized Wald test will be further illustrated in the examples presented in the Chap. 15.

Next two sections:

- GEE theory
- Use calculus and matrix notation

In the final two sections of this chapter we discuss the estimating equations used for GLM and GEE models. It is the estimating equations that form the underpinnings of a GEE analysis. The formulas presented use calculus and matrix notation for simplification. Although helpful, a background in these mathematical disciplines is not essential for an understanding of the material.

XI. Score Equations and “Score-like” Equations

L = likelihood function

ML solves estimating equations called *score equations*.

$$\left.\begin{aligned} S_1 &= \frac{\partial \ln L}{\partial \beta_0} = 0 \\ S_2 &= \frac{\partial \ln L}{\partial \beta_1} = 0 \\ &\cdot \\ &\cdot \\ &\cdot \\ S_{p+1} &= \frac{\partial \ln L}{\partial \beta_p} = 0 \end{aligned}\right\} \begin{array}{l} p+1 \text{ equations in} \\ p+1 \text{ unknowns} \\ (\beta\text{s}) \end{array}$$

The estimation of parameters often involves solving a system of equations called estimating equations. GLM utilizes maximum likelihood (ML) estimation methods. The likelihood is a function of the unknown parameters and the observed data. Once the likelihood is formulated, the parameters are estimated by finding the values of the parameters that maximize the likelihood. A common approach for maximizing the likelihood uses calculus. The partial derivatives of the log likelihood with respect to each parameter are set to zero. If there are $p + 1$ parameters, including the intercept, then there are $p + 1$ partial derivatives and, thus, $p + 1$ equations. These estimating equations are called *score equations*. The maximum likelihood estimates are then found by solving the system of score equations.

In GLM, score equations involve $\mu_i = E(Y_i)$ and $\text{var}(Y_i)$

For GLM, the score equations have a special form due to the fact that the responses follow a distribution from the exponential family. These score equations can be expressed in terms of the means (μ_i) and the variances [$\text{var}(Y_i)$] of the responses, which are modeled in terms of the unknown parameters (β_0, β_1, $\beta_2, \ldots, \beta_p$), and the observed data.

K = # of subjects

$p + 1$ = # of parameters (β_h, $h = 0, 1, 2, \ldots, p$)

Yields $p + 1$ score equations $S_1, S_2, \ldots, S_{p+1}$

(see formula on next page)

If there are K subjects, with each subject contributing one response, and $p + 1$ beta parameters (β_0, β_1, $\beta_2, \ldots, \beta_p$), then there are $p + 1$ score equations, one equation for each of the $p + 1$ beta parameters, with β_h being the ($h + 1$)st element of the vector of parameters.

$$S_{h+1} = \underbrace{\sum_{i=1}^{K} \frac{\partial\mu_i}{\beta_h}}_{\text{partial derivative}} \underbrace{[\text{var}(Y_i)]^{-1}}_{\text{variance}} \underbrace{[Y_i - \mu_i]}_{\text{residual}} = 0$$

Solution: iterative (by computer)

The $(h + 1)$st score equation (S_{h+1}) is written as shown on the left. For each score equation, the ith subject contributes a three-way product involving the partial derivative of μ_i with respect to a regression parameter, times the inverse of the variance of the response, times the difference between the response and its mean (μ_i).

The process of obtaining a solution to these equations is accomplished with the use of a computer and typically is iterative.

GLM score equations:

- Completely specified by $E(Y_i)$ and $\text{var}(Y_i)$
- Basis of QL estimation

A key property for GLM score equations is that they are completely specified by the mean and the variance of the random response. The entire distribution of the response is not really needed. This key property forms the basis of quasi-likelihood (QL) estimation.

QL estimation:

- **"Score-like" equations**
- No likelihood
- $\text{var}(Y_i) = \underbrace{\phi}_{\text{scale factor}} \underbrace{V(\mu_i)}_{\text{function of } \mu}$
- $g(\mu) = \beta_0 + \sum_{h=1}^{p} \beta_h X_h$
- solution yields QL estimates

Quasi-likelihood estimating equations follow the same form as score equations. For this reason, QL estimating equations are often called *"score-like" equations*. However, they are not score equations because the likelihood is not formulated. Instead, a relationship between the variance and mean is specified. The variance of the response, $\text{var}(Y_i)$, is set equal to a *scale factor* (ϕ) times a function of the mean response, $V(\mu_i)$. "Score-like" equations can be used in a similar manner as score equations in GLM. If the mean is modeled using a link function $g(\mu)$, QL estimates can be obtained by solving the system of "score-like" equations.

Logistic regression: $Y = (0, 1)$

$\mu = \text{P}(Y = 1|\mathbf{X})$

$V(\mu) = \text{P}(Y = 1|\mathbf{X})[1 - \text{P}(Y = 1|\mathbf{X})]$
$\quad = \mu(1 - \mu)$

For logistic regression, in which the outcome is coded 0 or 1, the mean response is the probability of obtaining the event, $\text{P}(Y = 1|\mathbf{X})$. The variance of the response equals $\text{P}(Y = 1|\mathbf{X})$ times 1 minus $\text{P}(Y = 1|\mathbf{X})$. So the relationship between the variance and mean can be expressed as $\text{var}(Y) = \phi\, V(\mu)$ where $V(\mu)$ equals μ times $(1 - \mu)$.

Scale factor = ϕ
Allows for *extra variation* in Y:

$\text{var}(Y) = \phi V(\mu)$

If Y binomial: $\phi = 1$ and $V(\mu) = \mu(1 - \mu)$

$\phi > 1$ indicates overdispersion
$\phi < 1$ indicates underdispersion

Equations	Allow extra variation?
QL: "score-like"	Yes
GLM: score	No

The scale factor ϕ allows for *extra variation* (dispersion) in the response beyond the assumed mean variance relationship of a binomial response, i.e., $\text{var}(Y) = \mu(1 - \mu)$. For the binomial distribution, the scale factor equals 1. If the scale factor is greater (or less) than 1, then there is overdispersion or underdispersion compared to a binomial response. The "score-like" equations are therefore designed to accommodate extra variation in the response, in contrast to the corresponding score equations from a GLM.

Summary: ML vs. QL Estimation

The process of ML and QL estimation can be summarized in a series of steps. These steps allow a comparison of the two approaches.

Step	ML Estimation	QL Estimation
1	Formulate L	–
2	For each β, obtain $\frac{\partial \ln \text{L}}{\partial \beta}$	–
3	Form score equations: $\left(\frac{\partial \ln L}{\partial \beta} = 0\right)$	Form "score-like" equations using $\text{var}(Y) = \phi V(\mu)$
4	Solve for ML estimates	Solve for QL estimates

ML estimation involves four steps:

Step 1. Formulate the likelihood in terms of the observed data and the unknown parameters from the assumed underlying distribution of the random data

Step 2. Obtain the partial derivatives of the log likelihood with respect to the unknown parameters

Step 3. Formulate score equations by setting the partial derivatives of the log likelihood to zero

Step 4. Solve the system of score equations to obtain the maximum likelihood estimates.

For QL estimation, the first two steps are bypassed by directly formulating and solving a system of "score-like" equations. These "score-like" equations are of a similar form as are the score equations derived for GLM. With GLM, the response follows a distribution from the exponential family, whereas with the "score-like" equations, the distribution of the response is not so restricted. In fact, the distribution of the response need not be known as long as the variance of the response can be expressed as a function of the mean.

XII. Generalizing the "Score-like" Equations to Form GEE Models

GEE models:

- For cluster-correlated data
- model parameters:

$$\underset{\substack{\nearrow\\ \text{regression}\\ \text{parameters}}}{\beta} \text{ and } \underset{\substack{\nwarrow\\ \text{correlation}\\ \text{parameters}}}{\alpha}$$

The estimating equations we have presented so far have assumed one response per subject. The estimating equations for GEE are "score-like" equations that can be used when there are several responses per subject or, more generally, when there are clustered data that contains within-cluster correlation. Besides the regression parameters (β) that are also present in a GLM, GEE models contain correlation parameters (α) to account for within-cluster correlation.

Matrix notation used to describe GEE

The most convenient way to describe GEE involves the use of matrices. Matrices are needed because there are several responses per subject and, correspondingly, a correlation structure to be considered. Representing these estimating equations in other ways becomes very complicated.

Matrices needed specific to each subject (cluster): $\mathbf{Y}_i$, $\boldsymbol{\mu}_i$, $\mathbf{D}_i$, $\mathbf{C}_i$, and $\mathbf{W}_i$

Matrices and vectors are indicated by the use of bold letters. The matrices that are needed are specific for each subject (i.e., ith subject), where each subject has n_i responses. The matrices are denoted as $\mathbf{Y}_i$, $\boldsymbol{\mu}_i$, $\mathbf{D}_i$, $\mathbf{C}_i$, and $\mathbf{W}_i$ and defined as follows:

$$\boldsymbol{Y}_i = \left\{ \begin{matrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{matrix} \right\} \text{ vector of } i\text{th subject's observed responses}$$

$\mathbf{Y}_i$ is the vector (i.e., collection) of the ith subject's observed responses.

$$\boldsymbol{\mu}_i = \left\{ \begin{matrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{in_i} \end{matrix} \right\} \text{ vector of } i\text{th subject's mean responses}$$

$\boldsymbol{\mu}_i$ is a vector of the ith subject's mean responses. The mean responses are modeled as functions of the predictor variables and the regression coefficients (as in GLM).

$\mathbf{C}_i$ = working correlation matrix ($n_i \times n_i$)

$\mathbf{C}_i$ is the $n_i \times n_i$ correlation matrix containing the correlation parameters. $\mathbf{C}_i$ is often referred to as the working correlation matrix.

$\mathbf{D_i}$ = diagonal matrix, with variance function $V(\mu_{ij})$ on diagonal

EXAMPLE

$n_i = 3$

$$\mathbf{D}_i = \begin{bmatrix} V(\mu_{i1}) & 0 & 0 \\ 0 & V(\mu_{i2}) & 0 \\ 0 & 0 & V(\mu_{i3}) \end{bmatrix}$$

$\mathbf{D}_i$ is a diagonal matrix whose jth diagonal (representing the jth observation of the ith subject) is the variance function $V(\mu_{ij})$. An example with three observations for subject i is shown at left. As a diagonal matrix, all the off-diagonal entries of the matrix are 0. Since $V(\mu_{ij})$ is a function of the mean, it is also a function of the predictors and the regression coefficients.

$\mathbf{W}_i$ = working covariance matrix ($n_i \times n_i$)

$$\mathbf{W}_i = \phi \mathbf{D}_i^{\frac{1}{2}} \mathbf{C}_i \mathbf{D}_i^{\frac{1}{2}}$$

$\mathbf{W}_i$ is an $n_i \times n_i$ variance–covariance matrix for the ith subjects’ responses, often referred to as the working covariance matrix. The variance–covariance matrix $\mathbf{W}_i$ can be decomposed into the scale factor (ϕ), times the square root of $\mathbf{D}_i$, times $\mathbf{C}_i$, times the square root of $\mathbf{D}_i$.

GEE: form similar to score equations

If K = # of subjects
n_i = # responses of subject i
$p + 1$ = # of parameters (β_h; $h = 0, 1, 2, \ldots, p$)

$$\text{GEE}_{h+1} = \sum_{i=1}^{K} \underbrace{\frac{\partial \boldsymbol{\mu}_i}{\beta_h}}_{\text{partial derivative}} \underbrace{[\mathbf{W}_i]^{-1}}_{\text{covariance}} \underbrace{[\mathbf{Y}_i - \boldsymbol{\mu}_i]}_{\text{residuel}} = 0$$

where

$$\mathbf{W}_i = \phi \mathbf{D}_i^{\frac{1}{2}} \mathbf{C}_i \mathbf{D}_i^{\frac{1}{2}}$$

Yields $p + 1$ GEE equations of the above form

The generalized estimating equations are of a similar form as the score equations presented in the previous section. If there are K subjects, with each subject contributing n_i responses, and $p + 1$ beta parameters ($\beta_0, \beta_1, \beta_2, \ldots, \beta_p$), with β_h being the ($h + 1$)st element of the vector of parameters, then the ($h + 1$)st estimating equation (GEE_{h+1}) is written as shown on the left.

There are $p + 1$ estimating equations, one equation for each of the $p + 1$ beta parameters. The summation is over the K subjects in the study. For each estimating equation, the ith subject contributes a three-way product involving the partial derivative of $\boldsymbol{\mu}_i$ with respect to a regression parameter, times the inverse of the subject’s variance–covariance matrix ($\mathbf{W}_i$), times the difference between the subject’s responses and their mean ($\boldsymbol{\mu}_i$).

Key difference GEE vs. GLM score equations: GEE allow for multiple responses per subject

The key difference between these estimating equations and the score equations presented in the previous section is that these estimating equations are generalized to allow for multiple responses from each subject rather than just one response. $\mathbf{Y}_i$ and $\boldsymbol{\mu}_i$ now represent a *collection* of responses (i.e., vectors) and $\mathbf{W}_i$ represents the variance–covariance matrix for all of the ith subject's responses.

GEE model parameters – three types:

There are three types of parameters in a GEE model. These are as follows.

1. **Regression parameters ($\boldsymbol{\beta}$)** Express relationship between predictors and outcome.

1. The *regression parameters* ($\boldsymbol{\beta}$) express the relationship between the predictors and the outcome. Typically, for epidemiological analyses, it is the regression parameters (or regression coefficients) that are of primary interest. The other parameters contribute to the accuracy and integrity of the model but are often considered "nuisance parameters". For a logistic regression, it is the regression parameter estimates that allow for the estimation of odds ratios.

2. **Correlation parameters ($\boldsymbol{\alpha}$)** Express within-cluster correlation; user specifies $\mathbf{C}_i$.

2. The *correlation parameters* ($\boldsymbol{\alpha}$) express the within-cluster correlation. To run a GEE model, the user specifies a correlation structure ($\mathbf{C}_i$), which provides a framework for the modeling of the correlation between responses from the same subject. The choice of correlation structure can affect both the estimates and the corresponding standard errors of the regression parameters.

3. **Scale factor** (ϕ) Accounts for extra variation of Y.

3. The *scale factor* (ϕ) accounts for overdispersion or underdispersion of the response. Overdispersion means that the data are showing more variation in the response variable than what is assumed from the modeling of the mean–variance relationship.

SLR: $\text{var}(Y) = \mu(1 - \mu)$

GEE logistic regression

$\text{var}(Y) = \phi\mu(1 - \mu)$

ϕ does not affect $\hat{\beta}$

ϕ affects $s_{\hat{\beta}}$ if $\phi \neq 1$

$\phi > 1$: overdispersion

$\phi < 1$: underdispersion

For a standard logistic regression (SLR), the variance of the response variable is assumed to be $\mu(1 - \mu)$, whereas for a GEE logistic regression, the variance of the response variable is modeled as $\phi\mu(1 - \mu)$ where ϕ is the scale factor. The scale factor does *not* affect the estimate of the regression parameters but it does affect their standard errors ($s_{\hat{\beta}}$) if the scale factor is different from 1. If the scale factor is greater than 1, there is an indication of overdispersion and the standard errors of the regression parameters are correspondingly scaled (inflated).

α and β estimated iteratively:

Estimates updated alternately

$\Rightarrow$ convergence

For a GEE model, the correlation parameters (α) are estimated by making use of updated estimates of the regression parameters (β), which are used to model the mean response. The regression parameter estimates are, in turn, updated using estimates of the correlation parameters. The computational process is iterative, by alternately updating the estimates of the alphas and then the betas until convergence is achieved.

To run GEE model, specify:

- $g(\mu)$ = link function
- $V(\mu)$ = mean variance relationship
- $\mathbf{C}_i$ = working correlation structure

GLM – no specification of a correlation structure

The GEE model is formulated by specifying a link function to model the mean response as a function of covariates (as in a GLM), a variance function which relates the mean and variance of each response, and a correlation structure that accounts for the correlation between responses within each cluster. For the user, the greatest difference of running a GEE model as opposed to a GLM is the specification of the correlation structure.

GEE logistic model:

$$\text{logit } P(D = 1|\mathbf{X}) = \beta_0 + \sum_{h=1}^{p} \beta_h X_h$$

α can affect estimation of β and $\sigma_{\hat{\beta}}$

but

$\hat{\beta}_i$ interpretation same as SLR

A GEE logistic regression is stated in a similar manner as a SLR, as shown on the left. The addition of the correlation parameters can affect the estimation of the beta parameters and their standard errors. However, the interpretation of the regression coefficients is the same as in SLR in terms of the way it reflects the association between the predictor variables and the outcome (i.e., the odds ratios).

GEE vs. Standard Logistic Regression

SLR equivalent to GEE model with:

1. Independent correlation structure
2. ϕ forced to equal 1
3. Model-based standard errors

With an SLR, there is an assumption that each observation is independent. By using an independent correlation structure, forcing the scale factor to equal 1, and using model-based rather than empirical standard errors for the regression parameter estimates, we can perform a GEE analysis and obtain results identical to those obtained from a standard logistic regression.

XIII. SUMMARY

✓ Chapter 14: Logistic Regression for Correlated Data: GEE

The presentation is now complete. We have described one analytic approach, the GEE model, for the situation where the outcome variable has dichotomous correlated responses. We examined the form and interpretation of the GEE model and discussed a variety of correlation structures that may be used in the formulation of the model. In addition, an overview of the mathematical theory underlying the GEE model has been presented.

We suggest that you review the material covered here by reading the detailed outline that follows. Then, do the practice exercises and test.

Chapter 15: GEE Examples

In the next chapter (Chap. 15), examples are presented to illustrate the effects of selecting different correlation structures for a model applied to a given dataset. The examples are also used to compare the GEE approach with a standard logistic regression approach in which the correlation between responses is ignored.

Detailed Outline

I. Overview (pages 492 – 493)

A. Focus: modeling outcomes with dichotomous correlated responses.

B. Observations can be subgrouped into clusters.

i. Assumption: responses are correlated within a cluster but independent between clusters.

ii. An analysis that ignores the within-cluster correlation may lead to incorrect inferences.

C. Primary analysis method examined is use of generalized estimating equations (GEE) model.

II. An example (Infant Care Study) (pages 493 – 498)

A. Example is a comparison of GEE to conventional logistic regression that ignores the correlation structure.

B. Ignoring the correlation structure can affect parameter estimates and their standard errors.

C. Interpretation of coefficients (i.e., calculation of odds ratios and confidence intervals) is the same as for standard logistic regression.

III. Data layout (page 499)

A. For repeated measures for K subjects:

i. The ith subject has n_i measurements recorded.

ii. The jth observation from the ith subject occurs at time t_{ij} with the outcome measured as Y_{ij} and with p covariates, $X_{ij1}, X_{ij2}, \ldots, X_{ijp}$.

B. Subjects do not have to have the same number of observations.

C. The time interval between measurements does not have to be constant.

D. The covariates may be time-independent or time-dependent for a given subject.

i. Time-dependent variable: values can vary between time intervals within a cluster;

ii. Time-independent variables: values do not vary between time intervals within a cluster.

IV. Covariance and correlation (pages 500 – 502)

A. Covariance of X and Y: the expected value of the product of X minus its mean and Y minus its mean:

$$\text{cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)].$$

B. Correlation: a standardized measure of covariance that is scale-free.

$$\rho_{xy} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

i. Correlation values range from -1 to $+1$.

ii. Can have correlations between observations on the same outcome variable.

iii. Can have correlations between dichotomous variables.

C. Correlation between observations in a cluster should be accounted for in the analysis.

V. **Generalized linear models** (pages 503 – 506)

A. Models in the class of GLM include logistic regression, linear regression, and Poisson regression.

B. Generalized linear model with p predictors is of the form

$$g(\mu) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i,$$

where μ is the mean response and $g(\mu)$ is a function of the mean

C. Three criteria for a GLM:

i. Random component: the outcome follows a distribution from the exponential family.

ii. Systematic component: the regression parameters are modeled linearly, as a function of the mean.

iii. Link function [$g(\mu)$]: this is the function that is modeled linearly with respect to the regression parameters:

a. Link function for logistic regression: logit function.

b. Inverse of link function [g^{-1} ($\mathbf{X}$, β)] $= \mu$.

c. For logistic regression, the inverse of the logit function is the familiar logistic model for the probability of an event:

$$g^{-1}(\mathbf{X}, \boldsymbol{\beta}) = \mu = \frac{1}{1 + \exp\left[-\left(\alpha + \sum_{i=1}^{p} \beta_i X_i\right)\right]}$$

D. GLM uses maximum likelihood methods for parameter estimation, which require specification of the full likelihood.

E. Quasi-likelihood methods provide an alternative approach to model development.

   i. A mean–variance relationship for the responses is specified.
   ii. The full likelihood is not specified.

VI. GEE models (pages 506–507)

A. GEE are generalizations of GLM.

B. In GEE models, as in GLM, a link function $[g(\mu)]$ is modeled as linear in the regression parameters.

   i. The logit link function is commonly used for dichotomous outcomes:

$$g(\mu) = \text{logit}[\text{P}(D = 1|\mathbf{X})] = \beta_0 + \sum_{h=1}^{p} \beta_h X_h$$

   ii. This model form is identical to the standard logistic model, but the underlying assumptions differ.

C. To apply a GEE model, a "working" correlation structure for within-cluster correlations is specified.

VII. Correlation structure (pages 507 – 510)

A. A correlation matrix in which responses are correlated within subjects and independent between subjects is in the form of a *block diagonal matrix*.

   i. Subject-specific matrices make up blocks along the diagonal.
   ii. All nondiagonal block entries are zero.

B. In a GEE model, each subject (cluster) has a common set of correlation parameters.

VIII. Different types of correlation structures (pages 511–516)

A. **Independent** correlation structure

   i. Assumption: responses within a cluster are uncorrelated.
   ii. The matrix for a given cluster is the identity matrix.

B. **Exchangeable** correlation structure

   i. Assumption: any two responses within a cluster have the same correlation (ρ).
   ii. Only one correlation parameter is estimated.
   iii. Therefore, the order of observations within a cluster is arbitrary.

C. **Autoregressive** correlation structure

   i. Often appropriate when there are repeated responses over time.

ii. The correlation is assumed to depend on the interval of time between responses.

iii. **AR1** is a special case of the autoregressive correlation structure:

a. Assumption of AR1: the correlation between any two responses from the same subject taken at time t_1 and t_2 is $\rho^{|t_1-t_2|}$.

b. There is one correlation parameter, but the order within a cluster is not arbitrary.

D. **Stationary *m*-dependent** correlation structure

i. Assumption: correlations k occasions apart are the same for $k = 1, 2, \ldots, m$, whereas correlations more than m occasions apart are zero.

ii. In a stationary m-dependent structure, there are m correlation parameters.

E. **Unstructured** correlation structure

i. In general, for n responses in a cluster, there are $n(n - 1)/2$ correlation parameters.

ii. Yields a separate correlation parameter for each pair ($j, k, j \neq k$) of observations within a cluster.

iii. The order of responses is not arbitrary.

F. **Fixed** correlation structure

i. The user specifies the values for the correlation parameters.

ii. No correlation parameters are estimated.

IX. Empirical and model-based variance estimators (pages 516–519)

A. If a GEE model is correctly specified (i.e., the correct link function and correlation structure are specified), the parameter estimates are consistent and the distribution of the estimates is asymptotically normal.

B. Even if the correlation structure is misspecified, the parameter estimates ($\hat{\beta}$) are consistent.

C. Two types of variance estimators can be obtained in GEE:

i. Model-based variance estimators.

a. Make use of the specified correlation structure.

b. Are consistent only if the correlation structure is correctly specified.

ii. Empirical (robust) estimators, which are an adjustment of model-based estimators:

a. Make use of the actual correlations between responses in the data as well as the specified correlation structure.

b. Are consistent even if the correlation structure is misspecified.

X. Statistical tests (pages 519–520)

A. **Score test**

i. The test statistic is based on the "score-like" equations.

ii. Under the null, the test statistic is distributed approximately chi-square with df equal to the number of parameters tested.

B. **Wald test**

i. For testing one parameter, the Wald test statistic is of the familiar form

$$Z = \frac{\hat{\beta}}{s_{\hat{\beta}}}.$$

ii. For testing more than one parameter, the *generalized Wald test* can be used.

iii. The generalized Wald test statistic is distributed approximately chi-square with df equal to the number of parameters approximate tested.

C. In GEE, the likelihood ratio test cannot be used because the likelihood is never formulated.

XI. Score equations and "score-like" equations (pages 521–523)

A. For maximum likelihood estimation, *score equations* are formulated by setting the partial derivatives of the log likelihood to zero for each unknown parameter.

B. In GLM, score equations can be expressed in terms of the means and variances of the responses.

i. Given $p+1$ beta parameters and β_h as the $(h+1)$st parameter, the $(h+1)$st score equation is

$$\sum_{i=1}^{K} \frac{\partial\mu_i}{\beta_h}[\mathrm{var}(Y_i)]^{-1}[Y_i - \mu_i] = 0,$$

where $h = 0, 1, 2, \ldots, p$.

ii. Note there are $p+1$ score equations, with summation over all K subjects.

C. Quasi-likelihood estimating equations follow the same form as score equations and thus are called *"score-like" equations*.

i. For quasi-likelihood methods, a mean variance relationship for the responses is specified [$V(\mu)$] but the likelihood in not formulated.

ii. For a dichotomous outcome with a binomial distribution, var(Y) = $\phi V(\mu)$, where $V(\mu) = \mu(1-\mu)$ and $\phi = 1$; in general ϕ is a scale factor that allows for extra variability in Y.

XII. Generalizing the "score-like" equations to form GEE models (pages 524 – 528)

A. GEE can be used to model clustered data that contains within cluster correlation.

B. Matrix notation is used to describe GEE:

i. $\mathbf{D}_i$ = diagonal matrix, with variance function $V(\mu_{ij})$ on diagonal.

ii. $\mathbf{C}_i$ = correlation matrix (or working correlation matrix).

iii. $\mathbf{W}_i$ = variance–covariance matrix (or working covariance matrix).

C. The form of GEE is similar to score equations:

$$\sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i'}{\beta_h}[\mathbf{W}_i]^{-1}[Y_i - \boldsymbol{\mu}_i] = 0,$$

where $\mathbf{W}_i = \phi \mathbf{D}_i^{\frac{1}{2}} \mathbf{C}_i \mathbf{D}_i^{\frac{1}{2}}$ and where $h = 0, 1, 2, \ldots, p$.

i. There are $p + 1$ estimating equations, with the summation over all K subjects.

ii. The key difference between generalized estimating equations and GLM score equations is that the GEE allow for multiple responses from each subject.

D. Three types of parameters in a GEE model:

i. Regression parameters ($\boldsymbol{\beta}$): these express the relationship between the predictors and the outcome. In logistic regression, the betas allow estimation of odds ratios.

ii. Correlation parameters ($\boldsymbol{\alpha}$): these express the within-cluster correlation. A working correlation structure is specified to run a GEE model.

iii. Scale factor (ϕ): this accounts for extra variation (underdispersion or overdispersion) of the response.

a. in a GEE logistic regression: $\text{var}(Y) = \phi\mu(1-\mu)$.

b. if different from 1, the scale factor (ϕ) will affect the estimated standard errors of the parameter estimates.

E. To formulate a GEE model, specify:

i. A link function to model the mean as a function of covariates.

ii. A function that relates the mean and variance of each response.

iii. a correlation structure to account for correlation between clusters.

F. Standard logistic regression is equivalent to a GEE logistic model with an independent correlation structure, the scale factor forced to equal 1, and model-based standard errors.

XIII. Summary (page 528)

Practice Exercises

Questions 1–5 pertain to identifying the following correlation structures that apply to clusters of four responses each:

$$\text{A} \quad \begin{bmatrix} 1 & 0.27 & 0.27 & 0.27 \\ 0.27 & 1 & 0.27 & 0.27 \\ 0.27 & 0.27 & 1 & 0..27 \\ 0.27 & 0.27 & 0.27 & 1 \end{bmatrix} \qquad \text{B} \quad \begin{bmatrix} 1 & 0.35 & 0 & 0 \\ 0.35 & 1 & 0.35 & 0 \\ 0 & 0.35 & 1 & 0.35 \\ 0 & 0 & 0.35 & 1 \end{bmatrix}$$

$$\text{C} \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \text{D} \quad \begin{bmatrix} 1 & 0.50 & 0.25 & 0.125 \\ 0.50 & 1 & 0.50 & 0.25 \\ 0.25 & 0.50 & 1 & 0.50 \\ 0.125 & 0.25 & 0.50 & 1 \end{bmatrix}$$

$$\text{E} \quad \begin{bmatrix} 1 & 0.50 & 0.25 & 0.125 \\ 0.50 & 1 & 0.31 & 0.46 \\ 0.25 & 0.31 & 1 & 0.163 \\ 0.125 & 0.46 & 0.163 & 1 \end{bmatrix}$$

1. Matrix A is an example of which correlation structure?
2. Matrix B is an example of which correlation structure?
3. Matrix C is an example of which correlation structure?
4. Matrix D is an example of which correlation structure?
5. Matrix E is an example of which correlation structure?

True or False (Circle T or F)

T F 6. If there are two responses for each cluster, then the exchangeable, AR1, and unstructured working correlation structure reduce to the same correlation structure.

T F 7. A likelihood ratio test can test the statistical significance of several parameters simultaneously in a GEE model.

T F 8. Since GEE models produce consistent estimates for the regression parameters even if the correlation structure is misspecified (assuming the mean response is modeled correctly), there is no particular advantage in specifying the correlation structure correctly.

T F 9. Maximum likelihood estimates are obtained in a GLM by solving a system of score equations.

The estimating equations used for GEE models have a similar structure to those score equations but are generalized to accommodate multiple responses from the same subject.

T F 10. If the correlation between X and Y is zero, then X and Y are independent.

Test

True or False (Circle T or F)

T F 1. It is typically the regression coefficients, not the correlation parameters, that are the parameters of primary interest in a correlated analysis.

T F 2. If an exchangeable correlation structure is specified in a GEE model, then the correlation between a subject's first and second responses is assumed equal to the correlation between the subject's first and third responses. However, that correlation can be different for each subject.

T F 3. If a dichotomous response, coded $Y = 0$ and $Y = 1$, follows a binomial distribution, then the mean response is the probability that $Y = 1$.

T F 4. In a GLM, the mean response is modeled as linear with respect to the regression parameters.

T F 5. In a GLM, a function of the mean response is modeled as linear with respect to the regression parameters. That function is called the link function.

T F 6. To run a GEE model, the user specifies a working correlation structure which provides a framework for the estimation of the correlation parameters.

T F 7. The decision as to whether to use model-based variance estimators or empirical variance estimators can affect both the estimation of the regression parameters and their standard errors.

T F 8. If a consistent estimator is used for a model, then the estimate should be correct even if the number of clusters is small.

T F 9. The empirical variance estimator allows for consistent estimation of the variance of the response variable even if the correlation structure is misspecified.

T F 10. Quasi-likelihood estimates may be obtained even if the distribution of the response variable is unknown. What should be specified is a function relating the variance to the mean response.

Answers to Practice Exercises

1. Exchangeable correlation structure
2. Stationary 1-dependent correlation structure
3. Independent correlation structure
4. Autoregressive (AR1) correlation structure
5. Unstructured correlation structure
6. T
7. F: the likelihood is never formulated in a GEE model
8. F: the estimation of parameters is more efficient [i.e., smaller var($\hat{\beta}$)] if the correct correlation structure is specified
9. T
10. F: the converse is true (i.e., if X and Y are independent, then the correlation is 0). The correlation is a measure of linearity. X and Y could have a nonlinear dependence and have a correlation of 0. In the special case where X and Y follow a normal distribution, then a correlation of 0 does imply independence.

15 GEE Examples

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_15, © Springer Science+Business Media, LLC 2010

Introduction

In this chapter, we present examples of GEE models applied to three datasets containing correlated responses. The examples demonstrate how to obtain odds ratios, construct confidence intervals, and perform statistical tests on the regression coefficients. The examples also illustrate the effect of selecting different correlation structures for a GEE model applied to the same data, and compare the results from the GEE approach with a standard logistic regression approach in which the correlation between responses is ignored.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

Objectives

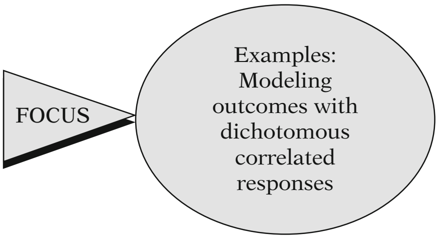
Upon completing this chapter, the learner should be able to:

1. State or recognize examples of correlated responses.
2. State or recognize when the use of correlated analysis techniques may be appropriate.
3. State or recognize examples of different correlation structures that may be used in a GEE model.
4. Given a printout of the results of a GEE model:

   i. State the formula and compute the estimated odds ratio
   ii. State the formula and compute a confidence interval for the odds ratio
   iii. Test hypotheses about the model parameters using the Wald test, generalized Wald test, or Score test, stating the null hypothesis and the distribution of the test statistic, and corresponding degrees of freedom under the null hypothesis
5. Recognize how running a GEE model differs from running a standard logistic regression on data with correlated dichotomous responses.
6. Recognize the similarities in obtaining and interpreting odds ratio estimates using a GEE model compared with a standard logistic regression model.

Presentation

I. Overview


FOCUS
Examples: Modeling outcomes with dichotomous correlated responses

In this chapter, we provide examples of how the GEE approach is used to carry out logistic regression for correlated dichotomous responses.

Three examples are presented:

1. Infant Care Study
2. Aspirin–Heart Bypass Study
3. Heartburn Relief Study

We examine a variety of GEE models using three databases obtained from the following studies: (1) Infant Care Study, (2) Aspirin–Heart Bypass Study, and (3) Heartburn Relief Study.

II. Example 1: Infant Care Study

Introduced in Chap. 14

In Chap. 14, we compared model output from two models run on data obtained from an infant care health intervention study in Brazil (Cannon et al., 2001). We continue to examine model output using these data, comparing the results of specifying different correlation structures.

Response (D):

Weight-for-height standardized (z) score

$$D = \begin{cases} 1 & \text{if } z < -1 \quad (\text{“Wasting”}) \\ 0 & \text{otherwise} \end{cases}$$

Recall that the outcome of interest is a dichotomous variable derived from a weight-for-height standardized score (i.e., z-score) obtained from the weight-for-height distribution of a reference population. The dichotomous outcome, an indication of “wasting,” is coded 1 if the z-score is less than negative 1, and 0 otherwise.

Independent variables:

BIRTHWGT (in grams)
GENDER

$$\text{DIARRHEA} = \begin{cases} 1 & \text{if symptoms present in past month} \\ 0 & \text{otherwise} \end{cases}$$

DIARRHEA

- Exposure of interest
- Time-dependent variable

The independent variables are BIRTHWGT (the weight in grams at birth), GENDER (1 = male, 2 = female), and DIARRHEA, a dichotomous variable indicating whether the infant had symptoms of diarrhea that month (1 = yes, 0 = no). We shall consider DIARRHEA as the main exposure of interest in this analysis. Measurements for each subject were obtained monthly for a 9-month period. The variables BIRTHWGT and GENDER are time-independent variables, as their values for a given individual do not change month to month. The variable DIARRHEA, however, is a time-dependent variable.

Infant Care Study Model

$$\begin{aligned} &\text{logit P}(D = 1|\mathbf{X}) \\ &\quad = \beta_0 + \beta_1\text{BIRTHWGT} \\ &\qquad + \beta_2\text{GENDER} + \beta_3\text{DIARRHEA} \end{aligned}$$

The model for the study can be stated as shown on the left.

Five GEE models presented, with different $\mathbf{C}_i$:

1. AR1 autoregressive
2. Exchangeable
3. Fixed
4. Independent
5. Independent (SLR)

Five GEE models are presented and compared, the last of which is equivalent to a standard logistic regression. The five models in terms of their correlation structure ($\mathbf{C}_i$) are as follows: (1) AR1 autoregressive, (2) exchangeable, (3) fixed, (4) independent, and (5) independent with model-based standard errors and scale factor fixed at a value of 1 [i.e., a standard logistic regression (SLR)]. After the output for all five models is shown, a table is presented that summarizes the results for the effect of the variable DIARRHEA on the outcome. Additionally, output from models using a stationary 4-dependent and a stationary 8-dependent correlation structure is presented in the Practice Exercises at the end of the chapter. A GEE model using an unstructured correlation structure did not converge for the Infant Care dataset using SAS version 9.2.

Output presented:

- $\hat{\beta}_h, s_{\hat{\beta}_h}$ (empirical), and Wald test P-values

Two sections of the output are presented for each model. The first contains the parameter estimate for each coefficient (i.e., beta), its estimated standard error (i.e., the square root of the estimated variance), and a P-value for the Wald test. Empirical standard errors rather than model-based are used for all but the last model. Recall that empirical variance estimators are consistent estimators even if the correlation structure is incorrectly specified (see Chap. 14).

- "Working" correlation matrix ($\mathbf{C}_i$) containing $\hat{\rho}$

The second section of output presented for each model is the working correlation matrix ($\mathbf{C}_i$). The working correlation matrix contains the estimates of the correlations, which depend on the specified correlation structure. The values of the correlation estimates are often not of primary interest. However, the examination of the fitted correlation matrices serves to illustrate key differences between the underlying assumptions about the correlation structure for these models.

Sample:

$K = 168$ infants, $n_i \leq 9$, but
9 infants "exposed cases":
(i.e., $D = 1$ and
DIARRHEA $= 1$ for any
month)

There are 168 clusters (infants) represented in the data. Only nine infants have a value of 1 for *both* the outcome and diarrhea variables at any time during their 9 months of measurements. The analysis, therefore, is strongly influenced by the small number of infants who are classified as "exposed cases" during the study period.

Model 1: AR1 correlation structure

```
                                 Empirical     Wald
Variable      Coefficient        Std Err      p-value

INTERCEPT       −1.3978          1.1960       0.2425
BIRTHWGT        −0.0005          0.0003       0.1080
GENDER           0.0024          0.5546       0.9965
DIARRHEA         0.2214          0.8558       0.7958
```

The parameter estimates for *Model 1* (autoregressive – AR1 correlation structure) are presented on the left. Odds ratio estimates are obtained and *interpreted* in a similar manner as in a standard logistic regression.

Effect of DIARRHEA:

$$\widehat{\text{OR}} = \exp(0.2214) = 1.25$$

$$\begin{aligned} 95\%\ \text{CI} &= \exp[0.2214 \pm 1.96(0.8558)] \\ &= (0.23, 6.68) \end{aligned}$$

For example, the estimated odds ratio for the effect of diarrhea symptoms on the outcome (a low weight-for-height *z*-score) is exp (0.2214) = 1.25. The 95% confidence interval can be calculated as exp[0.2214 ± 1.96 (0.8558)], yielding a confidence interval of (0.23, 6.68).

Working correlation matrix: 9 × 9

The working correlation matrix for each of these models contains nine rows and nine columns, representing an estimate for the month-to-month correlation between each infant's responses. Even though some infants did not contribute nine responses, the fact that each infant contributed *up to* nine responses accounts for the dimensions of the working correlation matrix.

AR1 working correlation matrix

(9 × 9 matrix: only three columns shown)

```
          COL1        COL2       ...       COL9

ROW1     1.0000      0.5254      ...      0.0058
ROW2     0.5254      1.0000      ...      0.0110
ROW3     0.2760      0.5254      ...      0.0210
ROW4     0.1450      0.2760      ...      0.0400
ROW5     0.0762      0.1450      ...      0.0762
ROW6     0.0400      0.0762      ...      0.1450
ROW7     0.0210      0.0400      ...      0.2760
ROW8     0.0110      0.0210      ...      0.5254
ROW9     0.0058      0.0110      ...      1.0000
```

The working correlation matrix for Model 1 is shown on the left. We present only columns 1, 2, and 9. However, all nine columns follow the same pattern.

The second-row, first-column entry of 0.5254 for the AR1 model is the estimate of the correlation between the first and second month measurements. Similarly, the third-row, first-column entry of 0.2760 is the estimate of the correlation between the first and third month measurements, which is assumed to be the same as the correlation between *any* two measurements that are 2 months apart (e.g., row 7, column 9). It is a property of the AR1 correlation structure that the correlation gets weaker as the measurements are further apart in time.

Estimated correlations:

$\hat{\rho} = 0.5254$ for responses 1 month apart (e.g., first and second)

$\hat{\rho} = 0.2760$ for responses 2 months apart (e.g., first and third, seventh and ninth)

$$\hat{\rho}_{j,j+1} = 0.5254$$
$$\hat{\rho}_{j,j+2} = (0.5254)^2 = 0.2760$$
$$\hat{\rho}_{j,j+3} = (0.5254)^3 = 0.1450$$

Note that the correlation between measurements 2 months apart (0.2760) is the square of measurements 1 month apart (0.5254), whereas the correlation between measurements 3 months apart (0.1450) is the cube of measurements 1 month apart. This is the key property of the AR1 correlation structure.

Model 2: Exchangeable correlation structure

Variable	Coefficient	Empirical Std Err	Wald *p*-value
INTERCEPT	−1.3987	1.2063	0.2463
BIRTHWGT	−0.0005	0.0003	0.1237
GENDER	−0.0262	0.5547	0.9623
DIARRHEA	**0.6485**	**0.7553**	**0.3906**

$\hat{\beta}_3$ for DIARRHEA = 0.6485 (vs. 0.2214 with Model 1)

Next we present the parameter estimates and working correlation matrix for a GEE model using the exchangeable correlation structure (*Model 2*). The coefficient estimate for DIARRHEA is 0.6485. This compares with the parameter estimate of 0.2214 for the same coefficient using the AR1 correlation structure in Model 1.

Exchangeable working correlation matrix

	COL1	COL2	...	COL9
ROW1	1.0000	0.4381	...	0.4381
ROW2	0.4381	1.0000	...	0.4381
ROW3	0.4381	0.4381	...	0.4381
ROW4	0.4381	0.4381	...	0.4381
ROW5	0.4381	0.4381	...	0.4381
ROW6	0.4381	0.4381	...	0.4381
ROW7	0.4381	0.4381	...	0.4381
ROW8	0.4381	0.4381	...	0.4381
ROW9	0.4381	0.4381	...	0.4381

Only one $\hat{\rho}$: $\hat{\rho} = 0.4381$

There is only one correlation to estimate with an exchangeable correlation structure. For this model, this estimate is 0.4381. The interpretation is that the correlation between any two outcome measures from the same infant is estimated at 0.4381 regardless of which months the measurements are taken.

Model 3: Fixed correlation structure

Variable	Coefficient	Empirical Std Err	Wald *p*-value
INTERCEPT	−1.3618	1.2009	0.2568
BIRTHWGT	−0.0005	0.0003	0.1110
GENDER	−0.0304	0.5457	0.9556
DIARRHEA	**0.2562**	**0.8210**	**0.7550**

Next we examine output from a model with a fixed, or user-defined, correlation structure (*Model 3*). The coefficient estimate and standard error for DIARRHEA are 0.2562 and 0.8210, respectively. These are similar to the estimates in the AR1 model, which were 0.2214 and 0.8558, respectively.

Fixed structure: ρ prespecified, not estimated

In Model 3, ρ fixed at 0.55 for consecutive months; 0.30 for nonconsecutive months.

Fixed working correlation matrix

```
         COL1       COL2      ...     COL9

ROW1    1.0000     0.5500     ...    0.3000
ROW2    0.5500     1.0000     ...    0.3000
ROW3    0.3000     0.5500     ...    0.3000
ROW4    0.3000     0.3000     ...    0.3000
ROW5    0.3000     0.3000     ...    0.3000
ROW6    0.3000     0.3000     ...    0.3000
ROW7    0.3000     0.3000     ...    0.3000
ROW8    0.3000     0.3000     ...    0.5500
ROW9    0.3000     0.3000     ...    1.0000
```

A fixed correlation structure has no correlation parameters to estimate. Rather, the values of the correlations are prespecified. For Model 3, the prespecified correlations are set at 0.55 between responses from consecutive months and 0.30 between responses from nonconsecutive months. For instance, the correlation between months 2 and 3 or months 2 and 1 is assumed to be 0.55, whereas the correlation between month 2 and the other months (not 1 or 3) is assumed to be 0.30.

Correlation structure (fixed) for Model 3: combines AR1 and exchangeable features

This particular selection of fixed correlation values contains some features of an autoregressive correlation structure, in that consecutive monthly measures are more strongly correlated. It also contains some features of an exchangeable correlation structure, in that, for nonconsecutive months, the order of measurements does not affect the correlation. Our choice of values for this model was influenced by the fitted values observed in the working correlation matrices of Model 1 and Model 2.

Choice of ρ at discretion of user, but may not always converge

The choice of correlation values for a fixed working correlation structure is at the discretion of the user. However, the parameter estimates are not guaranteed to converge for every choice of correlation values. In other words, the software package may not be able to provide parameter estimates for a GEE model for some user-defined correlation structures.

Allows flexibility specifying complicated $\mathbf{C}_i$

The use of a fixed correlation structure contrasts with other correlation structures in that the working correlation matrix ($\mathbf{C}_i$) does not result from fitting a model to the data, since the correlation values are all prespecified. However, it does allow flexibility in the specification of more complicated correlation patterns.

Independent correlation structure: two models

Model 4. Uses empirical $s_{\hat{\beta}}$;
ϕ not fixed

Model 5. Uses model-based $s_{\hat{\beta}}$;
ϕ fixed at 1

$S_{\hat{\beta}}$ affected

BUT

$\hat{\beta}$ *not* affected

Next, we examine output from models that incorporate an independent correlation structure (*Model 4* and *Model 5*). The key difference between Model 4 and a standard logistic regression (Model 5) is that Model 4 uses the empirical standard errors, whereas Model 5 uses the model-based standard errors. The other difference is that the scale factor is not preset equal to 1 in Model 4 as it is in Model 5. These differences only affect the standard errors of the regression coefficients rather than the estimates of the coefficients themselves.

Independent working correlation matrix

```
          COL1      COL2      ...     COL9

ROW1     1.0000    0.0000     ...    0.0000
ROW2     0.0000    1.0000     ...    0.0000
ROW3     0.0000    0.0000     ...    0.0000
ROW4     0.0000    0.0000     ...    0.0000
ROW5     0.0000    0.0000     ...    0.0000
ROW6     0.0000    0.0000     ...    0.0000
ROW7     0.0000    0.0000     ...    0.0000
ROW8     0.0000    0.0000     ...    0.0000
ROW9     0.0000    0.0000     ...    1.0000
```

Measurements on same subject assumed uncorrelated.

The working correlation matrix for an independent correlation structure is the identity matrix – with a 1 for the diagonal entries and a 0 for the other entries. The zeros indicate that the outcome measurements taken on the same subject are assumed uncorrelated.

Model 4: Independent correlation structure

```
                                         Wald
                         Empirical Std    p-
Variable    Coefficient       Err       value

INTERCEPT     −1.4362       1.2272     0.2419
BIRTHWGT      −0.0005       0.0003     0.1350
GENDER        −0.0453       0.5526     0.9346
DIARRHEA       0.7764       0.5857     0.1849
```

The outputs for Model 4 and Model 5 (next page) are shown on the left. The corresponding coefficients for each model are identical as expected. However, the estimated standard errors of the coefficients and the corresponding Wald test *P*-values differ for the two models.

Model 5: Standard logistic regression (naive model)

Variable	Coefficient	Model-based Std Err	Wald *p*-value
INTERCEPT	−1.4362	0.6022	0.0171
BIRTHWGT	−0.0005	0.0002	0.0051
GENDER	−0.0453	0.2757	0.8694
DIARRHEA	**0.7764**	**0.4538**	**0.0871**

$\hat{\beta}_3$ for DIARRHEA same *but* $s_{\hat{\beta}}$ and Wald *P*-values differ.

In particular, the coefficient estimate for DIARRHEA is 0.7764 in both Model 4 and Model 5; however, the standard error for DIARRHEA is larger in Model 4 at 0.5857 compared with 0.4538 for Model 5. Consequently, the *P*-values for the Wald test also differ: 0.1849 for Model 4 and 0.0871 for Model 5.

Model 4 vs. Model 5

- Parameter estimates same
- $s_{\hat{\beta}}$ Model 4 $>$ $s_{\hat{\beta}}$ Model 5

Other data: possible that
$s_{\hat{\beta}}$ (empirical) $<$ $s_{\hat{\beta}}$ (model based)

The other parameters in both models exhibit the same pattern, in that the coefficient estimates are the same, but the standard errors are larger for Model 4. In this example, the empirical standard errors are larger than their model-based counterparts, but this does not always occur. With other data, the reverse can occur.

Summary. Comparison of model results for DIARRHEA

Correlation structure	Odds ratio	95% CI
1 AR(1)	1.25	(0.23, 6.68)
2 Exchangeable	1.91	(0.44, 8.37)
3 Fixed (user defined)	1.29	(0.26, 6.46)
4 Independent	2.17	(0.69, 6.85)
5 Independent (SLR)	2.17	(0.89, 5.29)

A summary of the results for each model for the variable DIARRHEA is presented on the left. Note that the choice of correlation structure affects both the odds ratio estimates and the standard errors, which in turn affects the width of the confidence intervals. The largest odds ratio estimates are 2.17 from Model 4 and Model 5, which use an independent correlation structure. The 95% confidence intervals for all of the models are quite wide, with the tightest confidence interval (0.89, 5.29) occurring in Model 5, which is a standard logistic regression. The confidence intervals for the odds ratio for DIARRHEA include the null value of 1.0 for all five models.

Impact of misspecification

(usually) $\longrightarrow$ $s_{\hat{\beta}}$

$\longrightarrow$ $\widehat{\text{OR}}$

For Models 1–5:

$\widehat{\text{OR}}$ range $= 1.25-3.39$

Typically, a misspecification of the correlation structure has a stronger impact on the standard errors than on the odds ratio estimates. In this example, however, there is quite a bit of variation in the odds ratio estimates across the five models (from 1.25 for Model 1 to 2.17 for Model 4 and Model 5).

$\widehat{\text{OR}}$ range suggests model instability.

Instability likely due to small number (nine) of exposed cases.

This variation in odds ratio estimates suggests a degree of model instability and a need for cautious interpretation of results. Such evidence of instability may not have been apparent if only a single correlation structure had been examined. The reason the odds ratio varies as it does in this example is probably due to the relatively few infants who are exposed cases ($n = 9$) for any of their nine monthly measurements.

Which models to eliminate?

Models 4 and 5 (independent):
Evidence of correlated observations

It is easier to eliminate prospective models than to choose a definitive model. The working correlation matrices of the first two models presented (AR1 autoregressive and exchangeable) suggest that there is a positive correlation between responses for the outcome variable. Therefore, an independent correlation structure is probably not justified. This would eliminate Model 4 and Model 5 from consideration.

Model 2 (exchangeable):
If autocorrelation suspected

The exchangeable assumption for Model 2 may be less satisfactory in a longitudinal study if it is felt that there is autocorrelation in the responses. If so, that leaves Model 1 and Model 3 as the models of choice.

Remaining models: similar results:

Model 1 (AR1)
$\widehat{\text{OR}}(95\%\ \text{CI}) = 1.25(0.23, 6.68)$

Model 3 (fixed)
$\widehat{\text{OR}}\ (95\%\ \text{CI}) = 1.29(0.26, 6.46)$

Model 1 and Model 3 yield similar results, with an odds ratio and 95% confidence interval of 1.25 (0.23, 6.68) for Model 1 and 1.29 (0.26, 6.46) for Model 3. Recall that our choice of correlation values used in Model 3 was influenced by the working correlation matrices of Model 1 and Model 2.

III. Example 2: Aspirin–Heart Bypass Study

Data source: Gavaghan et al., 1991

Subjects: 214 patients received up to 6 coronary bypass grafts.

Randomly assigned to treatment group:

$$\text{ASPIRIN} = \begin{cases} 1 & \text{if daily aspirin} \\ 0 & \text{if daily placebo} \end{cases}$$

Response (D): Occlusion of a bypass graft 1 year later

$$D = \begin{cases} 1 & \text{if blocked} \\ 0 & \text{if unblocked} \end{cases}$$

Additional covariates:

AGE (in years)
GENDER (1 = male, 2 = female)
WEIGHT (in kilograms)
HEIGHT (in centimeters)

The next example uses data from a study in Sydney, Australia, which examined the efficacy of aspirin for prevention of thrombotic graft occlusion after coronary bypass grafting (Gavaghan et al., 1991). Patients ($K = 214$) were given a variable number of artery bypasses (up to six) in a single operation, and randomly assigned to take either aspirin (ASPIRIN = 1) or a placebo (ASPIRIN = 0) every day. One year later, angiograms were performed to check each bypass for occlusion (the outcome), which was classified as blocked ($D = 1$) or unblocked ($D = 0$). Additional covariates include AGE (in years), GENDER (1 = male, 2 = female), WEIGHT (in kilograms), and HEIGHT (in centimeters).

Correlation structures to consider:

- Exchangeable
- Independent

In this study, there is no meaningful distinction between artery bypass 1, artery bypass 2, or artery bypass 3 in the same subject. Since the order of responses within a cluster is arbitrary, we may consider using either the exchangeable or independent correlation structure. Other correlation structures make use of an inherent order for the within-cluster responses (e.g., monthly measurements), so they are not appropriate here.

Model 1: interaction model

Interaction terms between ASPIRIN and the other four covariates included.

$$\begin{aligned}
&\text{logit } P(D = 1|\mathbf{X}) \\
&= \beta_0 + \beta_1\text{ASPIRIN} + \beta_2\text{AGE} \\
&\quad + \beta_3\text{GENDER} + \beta_4\text{WEIGHT} \\
&\quad + \beta_5\text{HEIGHT} + \beta_6\text{ASPIRIN} \times \text{AGE} \\
&\quad + \beta_7\text{ASPIRIN} \times \text{GENDER} \\
&\quad + \beta_8\text{ASPIRIN} \times \text{WEIGHT} \\
&\quad + \beta_9\text{ASPIRIN} \times \text{HEIGHT}
\end{aligned}$$

The first model considered (*Model 1*) allows for interaction between ASPIRIN and each of the other four covariates. The model can be stated as shown on the left.

Exchangeable correlation structure

```
                                              Wald
                              Empirical Std    p-
Variable      Coefficient         Err        value

INTERCEPT       −1.1583          2.3950     0.6286
ASPIRIN          0.3934          3.2027     0.9022
AGE             −0.0104          0.0118     0.3777
GENDER          −0.9377          0.3216     0.0035
WEIGHT           0.0061          0.0088     0.4939
HEIGHT           0.0116          0.0151     0.4421
ASPIRIN          0.0069          0.0185     0.7087
   × AGE
ASPIRIN          0.9836          0.5848     0.0926
   × GENDER
ASPIRIN         −0.0147          0.0137     0.2848
   × WEIGHT
ASPIRIN         −0.0107          0.0218     0.6225
   × HEIGHT
```

Notice that the model contains a term for ASPIRIN, terms for the four covariates, and four product terms containing ASPIRIN. An exchangeable correlation structure is specified. The parameter estimates are shown on the left.

The output can be used to estimate the odds ratio for ASPIRIN = 1 vs. ASPIRIN = 0. If interaction is assumed, then a *different* odds ratio estimate is allowed for each pattern of covariates where the covariates interacting with ASPIRIN change values.

Odds ratio$_{\textbf{(ASPIRIN = 1 vs. ASPIRIN = 0)}}$

$$\begin{aligned}\text{odds} = \exp(&\beta_0 + \beta_1\text{ASPIRIN} + \beta_2\text{AGE}\\ &+ \beta_3\text{GENDER} + \beta_4\text{WEIGHT}\\ &+ \beta_5\text{HEIGHT} + \beta_6\text{ASPIRIN} \times \text{AGE}\\ &+ \beta_7\text{ASPIRIN} \times \text{GENDER}\\ &+ \beta_8\text{ASPIRIN} \times \text{WEIGHT}\\ &+ \beta_9\text{ASPIRIN} \times \text{HEIGHT})\end{aligned}$$

The odds ratio estimates can be obtained by separately inserting the values ASPIRIN = 1 and ASPIRIN = 0 in the expression of the odds shown on the left and then dividing one odds by the other.

Separate OR for each pattern of covariates:

$$\begin{aligned}\text{OR} = \exp(&\beta_1 + \beta_6\text{AGE} + \beta_7\text{GENDER}\\ &+ \beta_8\text{WEIGHT} + \beta_9\text{HEIGHT})\end{aligned}$$

This yields the expression for the *odds ratio*, also shown on the left.

AGE = 60, GENDER = 1, WEIGHT = 75 kg, HEIGHT = 170 cm

$$\begin{aligned}&\widehat{\text{OR}}_{\text{ASPIRIN = 1 vs. ASPIRIN = 0)}}\\ &\quad = \exp[0.3934 + (0.0069)(60)\\ &\qquad + (0.9836)(1) + (-0.0147)(75)\\ &\qquad + (-0.0107)(170)] = 0.32\end{aligned}$$

The odds ratio (comparing ASPIRIN status) for a 60-year-old male who weighs 75 kg and is 170 cm tall can be estimated using the output as 0.32.

Chunk test

H_0: $\beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$

A chunk test can be performed to determine if the four product terms can be dropped from the model. The null hypothesis is that the betas for the interaction terms are all equal to zero.

~~Likelihood ratio test~~

for GEE models

Recall for a standard logistic regression that the likelihood ratio test can be used to simultaneously test the statistical significance of several parameters. For GEE models, however, a likelihood is never formulated, which means that the likelihood ratio test cannot be used.

Two tests:

- Score test
- Generalized Wald test

Under H_0, both test statistics approximate χ^2 with
df = # of parameters tested.

There are two other statistical tests that can be utilized for GEE models. These are the generalized Score test and the generalized Wald test. The test statistic for the Score test relies on the "score-like" generalized estimating equations that are solved to produce the parameter estimates for the GEE model (see Chap. 14). The test statistic for the generalized Wald test generalizes the Wald test statistic for a single parameter by utilizing the variance–covariance matrix of the parameter estimates. The test statistics for both the Score test and the generalized Wald test follow an approximate chi-square distribution under the null with the degrees of freedom equal to the number of parameters that are tested.

Chunk test for interaction terms:

Type	DF	Chi-square	*P*-value
Score	4	3.66	0.4544
Wald	4	3.53	0.4737

Both tests fail to reject H_0.

The output for the Score test and the generalized Wald test for the four interaction terms is shown on the left. The test statistic for the Score test is 3.66 with the corresponding *p*-value at 0.45. The generalized Wald test yields similar results, as the test statistic is 3.53 with the *p*-value at 0.47. Both tests indicate that the null hypothesis should not be rejected and suggest that a model without the interaction terms may be appropriate.

Model 2: No interaction model (GEE)

$$\begin{aligned}\text{logit P}(D = 1|\mathbf{X}) &= \beta_0 + \beta_1\text{ASPIRIN} + \beta_2\text{AGE}\\ &\quad + \beta_3\text{GENDER} + \beta_4\text{WEIGHT}\\ &\quad + \beta_5\text{HEIGHT}\end{aligned}$$

The no interaction model (*Model 2*) is presented at left. The GEE parameter estimates using the exchangeable correlation structure are also shown.

Model 2 Output (Exchangeable)

```
                         Empirical     Wald
Variable    Coefficient   Std Err    p-value

INTERCEPT     −0.4713      1.6169     0.7707
ASPIRIN       −1.3302      0.1444     0.0001
AGE           −0.0086      0.0087     0.3231
GENDER        −0.5503      0.2559     0.0315
WEIGHT        −0.0007      0.0066     0.9200
HEIGHT         0.0080      0.0105     0.4448
```

Odds ratio

$$\widehat{\text{OR}}_{\text{ASPIRIN}=1\text{ vs. ASPIRIN}=0} = \exp(-1.3302) = 0.264$$

The odds ratio for aspirin use is estimated at exp(−1.3302) = 0.264, which suggests that aspirin is a preventive factor toward thrombotic graft occlusion after coronary bypass grafting.

Wald test

H_0: $\beta_1 = 0$

$$Z = \frac{-1.3302}{0.1444} = -9.21, \; P = 0.0001$$

The Wald test can be used for testing the hypothesis H_0: $\beta_1 = 0$. The value of the z test statistic is −9.21. The P-value of 0.0001 indicates that the coefficient for ASPIRIN is statistically significant.

Score test

H_0: $\beta_1 = 0$

Chi-square = 65.84, P = 0.0001

Alternatively, the Score test can be used to test the hypothesis H_0: $\beta_1 = 0$. The value of the chi-square test statistic is 65.34 yielding a similar statistically significant P-value of 0.0001.

Note:
$Z^2 = (-9.21)^2 = 84.82$,
so Wald ≠ Score

Note, however, that the χ^2 version of the Wald text (i.e., Z^2) differs from the Score statistic.

Exchangeable working correlation matrix

	COL1	COL2	...	COL6
ROW1	1.0000	−0.0954	...	−0.0954
ROW2	−0.0954	1.0000	...	−0.0954
ROW3	−0.0954	−0.0954	...	−0.0954
ROW4	−0.0954	−0.0954	...	−0.0954
ROW5	−0.0954	−0.0954	...	−0.0954
ROW6	−0.0954	−0.0954	...	1.0000

$\hat{\rho} = -0.0954$

The correlation parameter estimate obtained from the working correlation matrix is −0.0954, which suggests a negative association between reocclusion of different arteries from the same bypass patient compared with reocclusions from different patients.

Model 3: SLR (naive model)

Variable	Coefficient	Model-based Std Err	Wald p-value
INTERCEPT	−0.3741	2.0300	0.8538
ASPIRIN	−1.3410	0.1676	0.0001
AGE	−0.0090	0.0109	0.4108
GENDER	−0.5194	0.3036	0.0871
WEIGHT	−0.0013	0.0088	0.8819
HEIGHT	0.0078	0.0133	0.5580
SCALE	1.0000	0.0000	

The output for a standard logistic regression (SLR) is presented on the left for comparison with the corresponding GEE models. The parameter estimates for the standard logistic regression are similar to those obtained from the GEE model, although their standard errors are slightly larger.

Comparison of model results for ASPIRIN

Correlation structure	Odds ratio	95% CI
Exchangeable (GEE)	0.26	(0.20, 0.35)
Independent (SLR)	0.26	(0.19, 0.36)

A comparison of the odds ratio estimates with 95% confidence intervals for the no-interaction models of both the GEE model and SLR is shown on the left. The odds ratio estimates and 95% confidence intervals are very similar. This is not surprising, since only a modest amount of correlation is detected in the working correlation matrix ($\hat{\rho} = -0.0954$).

In this example, predictor values did not vary within a cluster.

In this example, none of the predictor variables (ASPIRIN, AGE, GENDER, WEIGHT, or HEIGHT) had values that varied within a cluster. This contrasts with the data used for the next example in which the exposure variable of interest is a time-dependent variable.

IV. Example 3: Heartburn Relief Study

Data source: Fictitious crossover study on heartburn relief.

Subjects: 40 patients; 2 symptom-provoking meals each; 1 of 2 treatments in random order

$$\text{Treatment (RX)} = \begin{cases} 1 & \text{if active RX} \\ 0 & \text{if standard RX} \end{cases}$$

Response (D): Relief from symptoms after 2 hours

$$D = \begin{cases} 1 & \text{if yes} \\ 0 & \text{if no} \end{cases}$$

The final dataset discussed is a fictitious crossover study on heartburn relief in which 40 subjects are given two symptom-provoking meals spaced a week apart. Each subject is administered an active treatment for heartburn (RX = 1) following one of the meals and a standard treatment (RX = 0) following the other meal in random order. The dichotomous outcome is relief from heartburn, determined from a questionnaire completed 2 hours after each meal.

Each subject has two observations

RX = 1

RX = 0

RX is time dependent: values change for each subject (cluster)

There are two observations recorded for each subject: one for the active treatment and the other for the standard treatment. The variable indicating treatment status (RX) is a time-dependent variable since it can change values within a cluster (subject). In fact, due to the design of the study, RX changes values in every cluster.

Model 1

$$\text{logit } P(D = 1|\mathbf{X}) = \beta_0 + \beta_1 \text{RX}$$

$n_i = 2$: [AR1, exchangeable, or unstructured]

$\Rightarrow$ same 2×2 $\mathbf{C}_i$

$$\mathbf{C}_i = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

For this analysis, RX is the only independent variable considered. The model is stated as shown on the left. With exactly two observations per subject, the only correlation to consider is the correlation between the two responses for the same subject. Thus, there is only one estimated correlation parameter, which is the same for each cluster. As a result, using an AR1, exchangeable, or unstructured correlation structure yields the same 2×2 working correlation matrix ($\mathbf{C}_i$).

Exchangeable correlation structure

Variable	Coefficient	Empirical Std Err	Wald *p*-value
INTERCEPT	−0.2007	0.3178	0.5278
RX	0.3008	0.3868	0.4368
Scale	1.0127	·	·

The output for a GEE model with an exchangeable correlation structure is presented on the left.

$$\widehat{\text{OR}} = \exp(0.3008) = 1.35$$

$$95\% \text{ CI} = (0.63, 2.88)$$

Exchangeable $\mathbf{C}_i$

	COL1	COL2
ROW1	1.0000	0.2634
ROW2	0.2634	1.0000

The odds ratio estimate for the effect of treatment for relieving heartburn is $\exp(0.3008) = 1.35$ with the 95% confidence interval of (0.63, 2.88). The working correlation matrix shows that the correlation between responses from the same subject is estimated at 0.2634.

SLR (naive) model

Variable	Coefficient	Model-based Std Err	Wald *p*-value
INTERCEPT	−0.2007	0.3178	0.5278
RX	0.3008	0.4486	0.5826
Scale	1.0000	·	·

$$\widehat{\text{OR}} = \exp(0.3008) = 1.35$$

$$95\% \text{ CI} = (0.56, 3.25)$$

A standard logistic regression is presented for comparison. The odds ratio estimate at $\exp(0.3008) = 1.35$ is exactly the same as was obtained from the GEE model with the exchangeable correlation structure; however, the standard error is larger, yielding a larger 95% confidence interval of (0.56, 3.25). Although an odds ratio of 1.35 suggests that the active treatment provides greater relief for heartburn, the null value of 1.00 is contained in the 95% confidence intervals for both models.

These examples illustrate the GEE approach for modeling data containing correlated dichotomous outcomes. However, use of the GEE approach is not restricted to dichotomous outcomes. As an extension of GLM, the GEE approach can be used to model other types of outcomes, such as count or continuous outcomes.

V. SUMMARY

✓ Chapter 15: GEE Examples

This presentation is now complete. The focus of the presentation was on several examples used to illustrate the application and interpretation of the GEE modeling approach. The examples show that the selection of different correlation structures for a GEE model applied to the same data can produce different estimates for regression parameters and their standard errors. In addition, we show that the application of a standard logistic regression model to data with correlated responses may lead to incorrect inferences.

We suggest that you review the material covered here by reading the detailed outline that follows. Then, do the practice exercises and test.

Chapter 16: Other Approaches to Analysis of Correlated Data

The GEE approach to correlated data has been used extensively. Other approaches to the analysis of correlated data are available. A brief overview of several of these approaches is presented in the next chapter.

Detailed Outline

I. Overview (page 542)

II–IV. Examples (pages 542–557)

A. Three examples were presented in detail:
   i. Infant Care Study
   ii. Aspirin–Heart Bypass Study
   iii. Heartburn Relief Study

B. Key points from the examples:
   i. The choice of correlation structure may affect both the coefficient estimate and the standard error of the estimate, although standard errors are more commonly impacted
   ii. Which correlation structure(s) should be specified depends on the underlying assumptions regarding the relationship between responses (e.g., ordering or time interval)
   iii. Interpretation of regression coefficients (in terms of odds ratios) is the same as in standard logistic regression

V. Summary (page 557)

Practice Exercises

The following printout summarizes the computer output from a GEE model run on the Infant Care Study data and should be used for Exercises 1–4. Recall that the data contained monthly information for each infant up to 9 months. The logit form of the model can be stated as follows:

$$\begin{aligned}\text{logit } P(X) = {} & \beta_0 + \beta_1\text{BIRTHWGT} + \beta_2\text{GENDER} \\ & + \beta_3\text{DIARRHEA}.\end{aligned}$$

The dichotomous outcome is derived from a weight-for-height z-score. The independent variables are BIRTHWGT (the weight in grams at birth), GENDER (1 = male, 2 = female), and DIARRHEA (a dichotomous variable indicating whether the infant had symptoms of diarrhea that month; coded 1 = yes, 0 = no).

A stationary 4-dependent correlation structure is specified for this model. Empirical and model-based standard errors are given for each regression parameter estimate. The working correlation matrix is also included in the output.

```
                              Empirical Std   Model-based
Variable   Coefficient             Err          Std Err

INTERCEPT    −2.0521            1.2323          0.8747
BIRTHWGT     −0.0005            0.0003          0.0002
GENDER        0.5514            0.5472          0.3744
DIARRHEA      0.1636            0.8722          0.2841
```

```
Stationary 4-Dependent Working Correlation Matrix

        COL1     COL2     COL3     COL4     COL5     COL6     COL7     COL8     COL9

ROW1  1.0000   0.5449   0.4353   0.4722   0.5334   0.0000   0.0000   0.0000   0.0000
ROW2  0.5449   1.0000   0.5449   0.4353   0.4722   0.5334   0.0000   0.0000   0.0000
ROW3  0.4353   0.5449   1.0000   0.5449   0.4353   0.4722   0.5334   0.0000   0.0000
ROW4  0.4722   0.4353   0.5449   1.0000   0.5449   0.4353   0.4722   0.5334   0.0000
ROW5  0.5334   0.4722   0.4353   0.5449   1.0000   0.5449   0.4353   0.4722   0.5334
ROW6  0.0000   0.5334   0.4722   0.4353   0.5449   1.0000   0.5449   0.4353   0.4722
ROW7  0.0000   0.0000   0.5334   0.4722   0.4353   0.5449   1.0000   0.5449   0.4353
ROW8  0.0000   0.0000   0.0000   0.5334   0.4722   0.4353   0.5449   1.0000   0.5449
ROW9  0.0000   0.0000   0.0000   0.0000   0.5334   0.4722   0.4353   0.5449   1.0000
```

1. Explain the underlying assumptions of a stationary 4-dependent correlation structure as it pertains to the Infant Care Study.
2. Estimate the odds ratio and 95% confidence interval for the variable DIARRHEA (1 vs. 0) on a low weight-for-height z-score (i.e., outcome = 1). Compute the 95% confidence interval in two ways: first using the empirical standard errors and then using the model-based standard errors.
3. Referring to Exercise 2: Explain the circumstances in which the model-based variance estimators yield consistent estimates.
4. Referring again to Exercise 2: Which estimate of the 95% confidence interval do you prefer?

The following output should be used for Exercises 5–10 and contains the results from running the same GEE model on the Infant Care data as in the previous questions, except that in this case, a stationary 8-dependent correlation structure is specified. The working correlation matrix for this model is included in the output.

```
                                          Empirical
Variable            Coefficient            Std Err

INTERCEPT             −1.4430              1.2084
BIRTHWGT              −0.0005              0.0003
GENDER                 0.0014              0.5418
DIARRHEA               0.3601              0.8122
```

```
Stationary 8-Dependent Working Correlation Matrix

        COL1     COL2     COL3     COL4     COL5     COL6     COL7     COL8     COL9

ROW1  1.0000   0.5255   0.3951   0.4367   0.4851   0.3514   0.3507   0.4346   0.5408
ROW2  0.5255   1.0000   0.5255   0.3951   0.4367   0.4851   0.3514   0.3507   0.4346
ROW3  0.3951   0.5255   1.0000   0.5255   0.3951   0.4367   0.4851   0.3514   0.3507
ROW4  0.4367   0.3951   0.5255   1.0000   0.5255   0.3951   0.4367   0.4851   0.3514
ROW5  0.4851   0.4367   0.3951   0.5255   1.0000   0.5255   0.3951   0.4367   0.4851
ROW6  0.3514   0.4851   0.4367   0.3951   0.5255   1.0000   0.5255   0.3951   0.4367
ROW7  0.3507   0.3514   0.4851   0.4367   0.3951   0.5255   1.0000   0.5255   0.3951
ROW8  0.4346   0.3507   0.3514   0.4851   0.4367   0.3951   0.5255   1.0000   0.5255
ROW9  0.5408   0.4346   0.3507   0.3514   0.4851   0.4367   0.3951   0.5255   1.0000
```

5. Compare the underlying assumptions of the stationary 8-dependent correlation structure with the unstructured correlation structure as it pertains to this model.
6. For the Infant Care data, how many more correlation parameters would be included in a model that uses an unstructured correlation structure rather than a stationary 8-dependent correlation structure?

7. How can the unstructured correlation structure be used to assess assumptions underlying other more constrained correlation structures?
8. Estimate the odds ratio and 95% confidence interval for DIARRHEA (1 vs. 0) using the model with the stationary 8-dependent working correlation structure.
9. If the GEE approach yields consistent estimates of the "true odds ratio" even if the correlation structure is misspecified, why are the odds ratio estimates different using a stationary 4-dependent correlation structure (Exercise 2) and a stationary 8-dependent correlation structure (Exercise 8).
10. Suppose that a parameter estimate obtained from running a GEE model on a correlated data set was not affected by the choice of correlation structure. Would the corresponding Wald test statistic also be unaffected by the choice of correlation structure?

Test

Questions 1–6 refer to models run on the data from the Heartburn Relief Study (discussed in Sect. IV). In that study, 40 subjects were given two symptom-provoking meals spaced a week apart. Each subject was administered an active treatment following one of the meals and a standard treatment following the other meal, in random order. The goal of the study was to compare the effects of an active treatment for heartburn with a standard treatment. The dichotomous outcome is relief from heartburn (coded 1 = yes, 0 = no). The exposure of interest is RX (coded 1 = active treatment, 0 = standard treatment). Additionally, it was hypothesized that the sequence in which each subject received the active and standard treatment could be related to the outcome. Moreover, it was speculated that the treatment sequence could be an effect modifier for the association between the treatment and heartburn relief. Consequently, two other variables are considered for the analysis: a dichotomous variable SEQUENCE and the product term RX*SEQ (RX times SEQUENCE). The variable SEQUENCE is coded 1 for subjects in which the active treatment was administered first and 0 for subjects in which the standard treatment was administered first.

The following printout summarizes the computer output for three GEE models run on the heartburn relief data (Model 1, Model 2, and Model 3). An exchangeable correlation structure is specified for each of these models. The variance–covariance matrix for the parameter estimates and the Score test for the variable RX*SEQ are included in the output for Model 1.

Model 1

```
                                           Empirical
Variable            Coefficient             Std Err

INTERCEPT             −0.6190                0.4688
RX                     0.4184                0.5885
SEQUENCE               0.8197                0.6495
RX*SEQ                −0.2136                0.7993

              Empirical Variance Covariance Matrix
                    For Parameter Estimates

              INTERCEPT      RX      SEQUENCE    RX*SEQ

INTERCEPT       0.2198   −0.1820   −0.2198     0.1820
RX             −0.1820    0.3463    0.1820    −0.3463
SEQUENCE       −0.2198    0.1820    0.4218    −0.3251
RX*SEQ          0.1820   −0.3463   −0.3251     0.6388

Score test statistic for RX*SEQ = 0.07
```

Model 2

```
                                        Empirical
 Variable          Coefficient           Std Err
INTERCEPT            −0.5625             0.4058
RX                    0.3104             0.3992
SEQUENCE              0.7118             0.5060
```

Model 3

```
                                        Empirical
 Variable          Coefficient           Std Err

INTERCEPT            −0.2007             0.3178
RX                    0.3008             0.3868
```

1. State the logit form of the model for Model 1, Model 2, and Model 3.
2. Use Model 1 to estimate the odds ratios and 95% confidence intervals for RX (active vs. standard treatment). *Hint.* Make use of the variance–covariance matrix for the parameter estimates.
3. In Model 1, what is the difference between the working covariance matrix and the covariance matrix for parameter estimates used to obtain the 95% confidence interval in the previous question?
4. Use Model 1 to perform the Wald test on the interaction term RX*SEQ at a 0.05 level of significance.
5. Use Model 1 to perform the Score test on the interaction term RX*SEQ at a 0.05 level of significance.
6. Estimate the odds ratio for RX using Model 2 and Model 3. Is there a suggestion that SEQUENCE is confounding the association between RX and heartburn relief. Answer this question from a data-based perspective (i.e., comparing the odds ratios) and a theoretical perspective (i.e., what it means to be a confounder).

Answers to Practice Exercises

1. The stationary 4-dependent working correlation structure uses four correlation parameters (α_1, α_2, α_3, and α_4). The correlation between responses from the same infant 1 month apart is α_1. The correlation between responses from the same infant 2, 3, or 4 months apart is α_2, α_3, and α_4, respectively. The correlation between responses from the same infant more than 4 months apart is assumed to be 0.
2. Estimated OR = exp(0.1636) = 1.18. 95% CI (with empirical SE): exp[0.1636 ± 1.96(0.8722)] = (0.21, 6.51); 95% CI (with model-based SE): exp[0.1636 ± 1.96(0.2841)] = (0.67, 2.06).
3. The model-based variance estimator would be a consistent estimator if the true correlation structure was stationary 4-dependent. In general, model-based variance estimators are more efficient $\{$i.e., smaller $\text{var}[\widehat{\text{var}}(\hat{\beta})]\}$ if the correlation structure is correctly specified.
4. The 95% confidence interval with the empirical standard errors is preferred since we cannot be confident that the true correlation structure is stationary 4-dependent.
5. The stationary 8-dependent correlation structure uses eight correlation parameters. With nine monthly responses per infant, each correlation parameter represents the correlation for a specific time interval between responses. The unstructured correlation structure, on the other hand, uses a different correlation parameter for each possible correlation for a given infant, yielding 36 correlation parameters. With the stationary 8-dependent correlation structure, the correlation between an infant's month 1 response and month 7 response is assumed to equal the correlation between an infant's month 2 response and month 8 response since the time interval between responses is the same (i.e., 6 months). The unstructured correlation structure does not make this assumption, using a different correlation parameter even if the time interval is the same.
6. There are $\frac{(9)(8)}{2} = 36$ correlation parameters using the unstructured correlation structure on the infant care data and 8 parameters using the stationary 8-dependent correlation structure. The difference is 28 correlation parameters.
7. By examining the correlation estimates in the unstructured working correlation matrix, we can evaluate which alternate, but more constrained, correlation structures seem reasonable. For example, if the

correlations are all similar, this would suggest that an exchangeable structure is reasonable.

8. Estimated OR = exp(0.3601) = 1.43.
   95% CI: exp[0.3601 ± 1.96(0.8122)] = (0.29, 7.04).
9. Consistency is an asymptotic property. As the number of clusters approaches infinity, the odds ratio estimate should approach the true odds ratio even if the correlation structure is misspecified. However, with a finite sample, the parameter estimate may still differ from the true parameter value. The fact that the parameter estimate for DIARRHEA is so sensitive to the choice of the working correlation structure demonstrates a degree of model instability.
10. No, because the Wald test statistic is a function of both the parameter estimate and its variance. Since the variance is typically affected by the choice of correlation structure, the Wald test statistic would also be affected.

16 Other Approaches for Analysis of Correlated Data

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_16, © Springer Science+Business Media, LLC 2010

Introduction

In this chapter, the discussion of methods to analyze outcome variables that have dichotomous correlated responses is expanded to include approaches other than GEE. Three other analytic approaches are discussed. These include the alternating logistic regressions algorithm, conditional logistic regression, and the generalized linear mixed model approach.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

Objectives

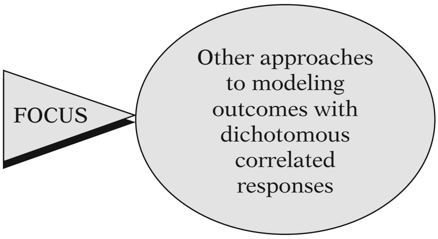
Upon completing this chapter, the learner should be able to:

1. Contrast the ALR method to GEE with respect to how within-cluster associations are modeled.
2. Recognize how a conditional logistic regression model can be used to handle subject-specific effects.
3. Recognize a generalized linear mixed (logistic) model.
4. Distinguish between random and fixed effects.
5. Contrast the interpretation of an odds ratio obtained from a marginal model with one obtained from a model containing subject-specific effects.

Presentation

I. Overview

In this chapter, we provide an introduction to modeling techniques other than GEE for use with dichotomous outcomes in which the responses are correlated.


FOCUS
Other approaches to modeling outcomes with dichotomous correlated responses

Other approaches for correlated data:

1. Alternating logistic regressions (ALR) algorithm
2. Conditional logistic regression
3. Generalized linear mixed model

In addition to the GEE approach, there are a number of alternative approaches that can be applied to model correlated data. These include (1) the alternating logistic regressions algorithm, which uses odds ratios instead of correlations, (2) conditional logistic regression, and (3) the generalized linear mixed model approach, which allows for random effects in addition to fixed effects. We briefly describe each of these approaches.

This chapter is not intended to provide a thorough exposition of these other approaches but rather an overview, along with illustrative examples, of other ways to handle the problem of analyzing correlated dichotomous responses. Some of the concepts that are introduced in this presentation are elaborated in the Practice Exercises at the end of the chapter.

Conditional logistic regression has previously been presented in Chap. 11 but is presented here in a some-what different context. The alternating logistic regression and generalized linear mixed model approaches for analyzing correlated dichotomous responses show great promise but at this point have not been fully investigated with regard to numerical estimation and possible biases.

II. The Alternating Logistic Regressions Algorithm

Modeling associations:

GEE approach	ALR approach
correlations (ρs)	odds ratios (ORs)

The alternating logistic regressions (ALR) algorithm is an analytic approach that can be used to model correlated data with dichotomous outcomes (Carey et al., 1993; Lipsitz et al., 1991). This approach is very similar to that of GEE. What distinguishes the two approaches is that with the GEE approach, associations between pairs of outcome measures are modeled with correlations, whereas with ALR, they are modeled with odds ratios. The odds ratio (OR_{ijk}) between the jth and kth responses for the ith subject can be expressed as shown on the left.

$$\mathrm{OR}_{ijk} = \frac{\mathrm{P}(Y_{ij}=1,\ Y_{ik}=1)\mathrm{P}(Y_{ij}=0,\ Y_{ik}=0)}{\mathrm{P}(Y_{ij}=1,\ Y_{ik}=0)\mathrm{P}(Y_{ij}=0,\ Y_{ik}=1)}$$

GEE: αs and βs estimated by alternately updating estimates until convergence

Recall that in a GEE model, the correlation parameters (α) are estimated using estimates of the regression parameters (β). The regression parameter estimates are, in turn, updated using estimates of the correlation parameters. The computational process alternately updates the estimates of the alphas and then the betas until convergence is achieved.

ALR: αs and βs estimated similarly

BUT

ALR : α are log ORs

(GEE : α are ρs)

The ALR approach works in a similar manner, except that the alpha parameters are log odds ratio parameters rather than correlation parameters. Moreover, for the same data, an odds ratio between the jth and kth responses that is greater than 1 using an ALR model corresponds to a positive correlation between the jth and kth responses using a GEE model. Similarly, an odds ratio less than 1 using an ALR model corresponds to a negative correlation between responses.

$$\text{ALR } \mathrm{OR}_{jk} > 1 \Leftrightarrow \text{GEE } \rho_{jk} > 1$$

$$\text{ALR } \mathrm{OR}_{jk} < 1 \Leftrightarrow \text{GEE } \rho_{jk} < 1$$

Same OR can correspond to different ρs

$$\begin{tikzcd}
 & \rho_{a_{jk}} \\
\mathrm{OR}_{jk} \arrow[ur] \arrow[dr] & \\
 & \rho_{b_{jk}}
\end{tikzcd}$$

However, the correspondence is not one-to-one, and examples can be constructed in which the same odds ratio corresponds to different correlations (see Practice Exercises 1–3).

ALR: dichotomous outcomes only
GEE: dichotomous and other outcomes are allowed

For many health scientists, an odds ratio measure, such as that provided with an ALR model, is more familiar and easier to interpret than a correlation measure. However, ALR models can only be used if the outcome is a dichotomous variable. In contrast, GEE models are not so restrictive.

EXAMPLE

GEE vs. ALR

Aspirin–Heart Bypass Study (Gavaghan et al., 1991)

The ALR model is illustrated by returning to the Aspirin–Heart Bypass Study example, which was first presented in Chap. 15. Recall that in that study, researchers examined the efficacy of aspirin for prevention of thrombotic graft occlusion after coronary bypass grafting in a sample of 214 patients (Gavaghan et al., 1991).

Subjects: received up to six coronary bypass grafts

Randomly assigned to treatment group:

$$\text{ASPIRIN} = \begin{cases} 1 & \text{if daily aspirin} \\ 0 & \text{if daily placebo} \end{cases}$$

Response (D): occlusion of a bypass graft 1 year later.

$$D = \begin{cases} 1 & \text{if blocked} \\ 0 & \text{if unblocked} \end{cases}$$

Additional covariates:

AGE (in years)
GENDER (1 = male, 2 = female)
WEIGHT (in kilograms)
HEIGHT (in centimeters)

Patients were given a variable number of artery bypasses (up to six) and randomly assigned to take either aspirin (ASPIRIN = 1) or a placebo (ASPIRIN = 0) every day. One year later, each bypass was checked for occlusion and the outcome was coded as blocked ($D = 1$) or unblocked ($D = 0$). Additional covariates included AGE (in years), GENDER (1 = male, 2 = female), WEIGHT (in kilograms), and HEIGHT (in centimeters).

Model:

$$\text{logit } P(\mathbf{X}) = \beta_0 + \beta_1\text{ASPIRIN} + \beta_2\text{AGE} + \beta_3\text{GENDER} + \beta_4\text{WEIGHT} + \beta_5\text{HEIGHT}$$

Consider the model presented at left, with ASPIRIN, AGE, GENDER, WEIGHT, and HEIGHT as covariates.

EXAMPLE (continued)

GEE Approach (Exchangeable ρ)

```
                      Empirical   z Wald
Variable  Coefficient  Std Err   p-value

INTERCEPT −0.4713     1.6169     0.7707
ASPIRIN   −1.3302     0.1444     0.0001
AGE       −0.0086     0.0087     0.3231
GENDER    −0.5503     0.2559     0.0315
WEIGHT    −0.0007     0.0066     0.9200
HEIGHT     0.0080     0.0105     0.4448
Scale      1.0076
```

Output from using the GEE approach is presented on the left. An exchangeable correlation structure is assumed. (This GEE output has previously been presented in Chap. 15)

Exchangeable $\mathbf{C}_i$ (GEE: $\hat{\rho} = -0.0954$)

```
         COL1      COL2    · · ·  COL6

ROW1     1.0000   -0.0954  · · · -0.0954
ROW2    -0.0954    1.0000  · · · -0.0954
ROW3    -0.0954   -0.0954  · · · -0.0954
ROW4    -0.0954   -0.0954  · · · -0.0954
ROW5    -0.0954   -0.0954  · · · -0.0954
ROW6    -0.0954   -0.0954  · · ·  1.0000
```

The correlation parameter estimate obtained from the working correlation matrix of the GEE model is −0.0954, which suggests a negative association between reocclusions on the same bypass patient.

ALR approach (Exchangeable OR)

```
                       Empirical   z wald
Variable  Coefficient   Std Err   p-value

INTERCEPT   −0.4806     1.6738     0.7740
ASPIRIN     −1.3253     0.1444     0.0001
AGE         −0.0086     0.0088     0.3311
GENDER      −0.5741     0.2572     0.0256
WEIGHT      −0.0003     0.0066     0.9665
HEIGHT       0.0077     0.0108     0.4761
ALPHA1      −0.4716     0.1217     0.0001
```

$\text{exp(ALPHA1)} = \widehat{\text{OR}}_{jk}\text{(exchangeable)}$

Output obtained from SAS PROC GENMOD using the ALR approach is shown on the left for comparison. An exchangeable *odds ratio* structure is assumed. The assumption underlying the exchangeable odds ratio structure is that the odds ratio between the ith subject's jth and kth responses is the same (for all j and k, $j \neq k$). The estimated exchangeable odds ratio is obtained by exponentiating the coefficient labeled ALPHA1.

EXAMPLE (continued)

Odds ratios

$\widehat{\mathbf{OR}}_{\text{ASPIRIN}=1 \text{ vs. } \text{ASPIRIN}=0}$:

GEE → $\exp(-1.3302) = 0.264$

ALR → $\exp(-1.3253) = 0.266$

S.E. (Aspirin) = 0.1444 (GEE and ALR)

The regression parameter estimates are very similar for the two models. The odds ratio for aspirin use on artery reocclusion is estimated as $\exp(-1.3302) = 0.264$ using the GEE model and $\exp(-1.3253) = 0.266$ using the ALR model. The standard errors for the aspirin parameter estimates are the same in both models (0.1444), although the standard errors for some of the other parameters are slightly larger in the ALR model.

Measure of association ($\widehat{\mathbf{OR}}_{jk}$)

$$\widehat{\text{OR}}_{jk} = \exp(\text{ALPHA1}) = \exp(-0.4716) = 0.62$$

(Negative association: similar to $\hat{\rho} = -0.0954$

The corresponding measure of association (the odds ratio) estimate from the ALR model can be found by exponentiating the coefficient of ALPHA1. This odds ratio estimate is $\exp(-0.4716) = 0.62$. As with the estimated exchangeable correlation ($\hat{\rho}$) from the GEE approach, the exchangeable OR estimate, which is less than 1, also indicates a negative association between any pair of outcomes (i.e., reocclusions on the same bypass patient).

95% CI for ALPHA1

$$= \exp[(-0.4716 \pm 1.96(0.1217)] = (0.49, 0.79)$$

P-value = 0.0001
⇒ ALPHA1 significant

A 95% confidence interval for the OR can be calculated as $\exp[-0.4716 \pm 1.96(0.1217)]$, which yields the confidence interval (0.49, 0.79). The P-value for the Wald test is also given in the output at 0.0001, indicating the statistical significance of the ALPHA1 parameter.

	GEE (ρ)	ALR (ALPHA1)
SE?	No	Yes
Test?	No	Yes

For the GEE model output, an estimated standard error (SE) or statistical test is not given for the correlation estimate. This is in contrast to the ALR output, which provides a standard error and statistical test for ALPHA1.

Key difference: GEE vs. ALR

GEE: ρ_{jk} are typically nuisance parameters

ALR: OR_{jk} are parameters of interest

This points out a key difference in the GEE and ALR approaches. With the GEE approach, the correlation parameters are typically considered to be nuisance parameters, with the parameters of interest being the regression coefficients (e.g., ASPIRIN). In contrast, with the ALR approach, the association between different responses is also considered to be of interest. Thus, the ALR approach allows statistical inferences to be assessed from both the alpha parameter and the beta parameters (regression coefficients).

ALR: allows inferences about both $\hat{\alpha}$ *and* $\hat{\beta}$s

III. Conditional Logistic Regression

EXAMPLE

Heartburn Relief Study
(“subject” as matching factor)

40 subjects received:

- Active treatment (“exposed”)
- Standard treatment (“unexposed”)

CLR model

$$\text{logit } P(\mathbf{X}) = \beta_0 + \beta_1\text{RX} + \sum_{i=1}^{39} \gamma_i V_i,$$

where

$$V_i = \begin{cases} 1 & \text{for subject } i \\ 0 & \text{otherwise} \end{cases}$$

GEE model

$$\text{logit } P(\mathbf{X}) = \beta_0 + \beta_1\text{RX}$$

CLR	**vs.**	**GEE**
↓		↓
39 V_i (dummy variables)		no V_i

Another approach that is applicable for certain types of correlated data is a matched analysis. This method can be applied to the Heartburn Relief Study example, with “subject” used as the matching factor. This example was presented in detail in Chap. 15. Recall that the dataset contained 40 subjects, each receiving an active or standard treatment for the relief of heartburn. In this framework, within each matched stratum (i.e., subject), there is an exposed observation (the active treatment) and an unexposed observation (the standard treatment). A conditional logistic regression (CLR) model, as discussed in Chap. 11, can then be formulated to perform a matched analysis. The model is shown on the left.

This model differs from the GEE model for the same data, also shown on the left, in that the conditional model contains 39 dummy variables besides RX. Each of the parameters (γ_i) for the 39 dummy variables represents the (fixed) effects for each of 39 subjects on the outcome. The 40th subject acts as the reference group since all of the dummy variables have a value of zero for the 40th subject (see Chap. 11).

CLR approach ⇒
responses assumed independent

Subject-specific γ_i allows for conditioning by subject

fixed effect

Responses can be independent if conditioned by subject

When using the CLR approach for modeling P(**X**), the responses from a specific subject are assumed to be independent. This may seem surprising since throughout this chapter we have viewed two or more responses on the same subject as likely to be correlated. Nevertheless, when dummy variables are used for each subject, each subject has his/her own subject-specific fixed effect included in the model. The addition of these subject-specific fixed effects can account for correlation that may exist between responses from the same subject in a GEE model. In other words, responses can be independent if *conditioned* by subject. However, this is not always the case. For example, if the actual underlying correlation structure is autoregressive, conditioning by subject would not account for the within-subject autocorrelation.

EXAMPLE (continued)

Model 1: conditional logistic regression

Variable	Coefficient	Std. error	Wald *P*-value
RX	0.4055	0.5271	0.4417

Returning to the Heartburn Relief Study data, the output obtained from running the conditional logistic regression is presented on the left.

No β_0 or γ_i estimates in CLR model (cancel out in conditional likelihood)

With a conditional logistic regression, parameter estimates are not obtained for the intercept or the dummy variables representing the matched factor (i.e., subject). These parameters cancel out in the expression for the conditional likelihood. However, this is not a problem because the parameter of interest is the coefficient of the treatment variable (RX).

Odds ratio and 95% CI

$\widehat{\text{OR}} = \exp(0.4055) = 1.50$

95% CI $= (0.534,\ 4.214)$

The odds ratio estimate for the effect of treatment for relieving heartburn is exp (0.4055) = 1.50, with a 95% confidence interval of (0.534, 4.214).

EXAMPLE (continued)

Model comparison

Model	OR	$s_{\hat{\beta}}$
CLR	1.50	0.5271
GEE	1.35	0.3868
SLR	1.35	0.4486

The estimated odds ratios and the standard errors for the parameter estimate for RX are shown at left for the conditional logistic regression (CLR) model, as well as for the GEE and standard logistic regression (SLR) discussed in Chap. 15. The odds ratio estimate for the CLR model is somewhat larger than the estimate obtained at 1.35 using the GEE approach. The standard error for the RX coefficient estimate in the CLR model is also larger than what was obtained in either the GEE model using empirical standard errors or in the standard logistic regression, which uses model-based standard errors.

	Estimation of predictors	
Analysis	Within-subject variability	Between-subject variability
Matched (CLR)	√	
Correlated (GEE)	√	√

No within-subject variability for an independent variable

$\Downarrow$

parameter will not be estimated using CLR

An important distinction between the CLR and GEE analytic approaches concerns the treatment of the predictor (independent) variables in the analysis. A matched analysis (CLR) relies on within-subject variability (i.e., variability within the matched strata) for the estimation of its parameters. A correlated (GEE) analysis takes into account both within-subject variability and between-subject variability. In fact, if there is no within-subject variability for an independent variable (e.g., a time-independent variable), then its coefficient cannot be estimated using a conditional logistic regression. In that situation, the parameter cancels out in the expression for the conditional likelihood. This is what occurs to the intercept as well as to the coefficients of the matching factor dummy variables when CLR is used.

EXAMPLE

CLR with time-independent predictors (Aspirin–Heart Bypass Study)

Subjects: 214 patients received up to 6 coronary bypass grafts.

Treatment:

$$\text{ASPIRIN} = \begin{cases} 1 & \text{if daily aspirin} \\ 0 & \text{if daily placebo} \end{cases}$$

$$D = \begin{cases} 1 & \text{if graft blocked} \\ 0 & \text{if graft unblocked} \end{cases}$$

To illustrate the consequences of only including independent variables with no within-cluster variability in a CLR, we return to the Aspirin–Heart Bypass Study discussed in the previous section. Recall that patients were given a variable number of artery bypasses in a single operation and randomly assigned to either aspirin or placebo therapy. One year later, angiograms were performed to check each bypass for reocclusion.

EXAMPLE (continued)

$$\begin{aligned}\text{logit P}(\mathbf{X}) = {} & \beta_0 + \beta_1\text{ASPIRIN} + \beta_2\text{AGE} \\ & + \beta_3\text{GENDER} \\ & + \beta_4\text{WEIGHT} \\ & + \beta_5\text{HEIGHT} + \sum_{i=1}^{213} \gamma_i V_i\end{aligned}$$

CLR model

```
                              Standard     Wald
Variable  Coefficient          Error     p-value

AGE              0               .          .
GENDER           0               .          .
WEIGHT           0               .          .
HEIGHT           0               .          .
ASPIRIN          0               .          .
```

Besides ASPIRIN, additional covariates include AGE, GENDER, WEIGHT, and HEIGHT. We restate the model from the previous section at left, which also includes 213 dummy variables for the 214 study subjects.

The output from running a conditional logistic regression is presented on the left. Notice that all of the coefficient estimates are zero with their standard errors missing. This indicates that the model did not execute. The problem occurred because none of the independent variables changed their values within any cluster (subject). In this situation, *all* of the predictor variables are said to be *concordant* in all the matching strata and uninformative with respect to a matched analysis. Thus, the conditional logistic regression, in effect, discards all of the data.

All strata
concordant ⇒ model will not run

Within-subject variability for one or more independent variable

⇓

- Model will run
- Parameters estimated for only those variables

If at least one variable in the model does vary within a cluster (e.g., a time-dependent variable), then the model will run. However, estimated coefficients will be obtained only for those variables that have within-cluster variability.

Matched analysis:

- Advantage: control of confounding factors
- Disadvantage: cannot separate effects of time-independent factors

An advantage of using a matched analysis with *subject* as the matching factor is the ability to control for potential confounding factors that can be difficult or impossible to measure. When the study subject is the matched variable, as in the Heartburn Relief example, there is an implicit control of fixed genetic and environmental factors that comprise each subject. On the other hand, as the Aspirin–Heart bypass example illustrates, a disadvantage of this approach is that we cannot model the separate effects of fixed time-independent factors. In this analysis, we cannot examine the separate effects of aspirin use, gender, and height using a matched analysis, because the values for these variables do not vary for a given subject.

Heartburn Relief Model:

(Subject modeled as *fixed effect*)

$$\text{logit P}(\mathbf{X}) = \beta_0 + \beta_1\text{RX} + \sum_{i=1}^{39} \gamma_i V_i,$$

where

$$V_i = \begin{cases} 1 & \text{for subject } i \\ 0 & \text{otherwise} \end{cases}$$

Alternative approach:
Subject modeled as *random effect*

What if study is replicated?

Different sample
⇒ different subjects
β_1 unchanged (fixed effect)
γ different

Parameters themselves may be random (not just their estimates)

With the conditional logistic regression approach, *subject* is modeled as a *fixed* effect with the gamma parameters (γ), as shown on the left for the Heartburn Relief example.

An alternative approach is to model subject as a *random* effect.

To illustrate this concept, suppose we attempted to replicate the heartburn relief study using a different sample of 40 subjects. We might expect the estimate for β_1, the coefficient for RX, to change due to sampling variability. However, the true value of β_1 would remain unchanged (i.e., β_1 is a fixed effect). In contrast, because there are different subjects in the replicated study, the parameters representing subject (i.e., the gammas) would therefore also be different. This leads to an additional source of variability that is not considered in the CLR, in that some of the parameters themselves (and not just their estimates) are random.

In the next section, we present an approach for modeling *subject* as a random effect, which takes into account that the subjects represent a random sample from a larger population.

IV. The Generalized Linear Mixed Model Approach

Mixed models:

- Random effects
- Fixed effects
- Cluster effect is random variable

The generalized linear mixed model (GLMM) provides another approach that can be used for correlated dichotomous outcomes. GLMM is a generalization of the linear mixed model. Mixed models refer to the *mixing* of random and fixed effects. With this approach, the cluster variable is considered a random effect. This means that the cluster effect is a random variable following a specified distribution (typically a normal distribution).

Mixed logistic model (MLM):

- Special case of GLMM
- Combines GEE and CLR features

GEE	CLR
User specifies $g(\mu)$ and $\mathbf{C}_i$	Subject-specific effects

GLMM: subject-specific effects *random*

A special case of the GLMM is the mixed logistic model (MLM). This type of model combines some of the features of the GEE approach and some of the features of the conditional logistic regression approach. As with the GEE approach, the user specifies the logit link function and a structure ($\mathbf{C}_i$) for modeling response correlation. As with the conditional logistic regression approach, subject-specific effects are directly included in the model. However, here these subject-specific effects are treated as random rather than fixed effects. The model is commonly stated in terms of the ith subject's mean response (μ_i).

EXAMPLE

Heartburn Relief Study

$\text{logit } \mu_i = \beta_0 + \beta_1 \text{RX}_i + b_{0i}$

β_1 = fixed effect

b_{0i} = random effect,

where b_{0i} is a random variable $\sim N(0, \sigma_{b_0}{}^2)$

We again use the heartburn data to illustrate the model (shown on the left) and state it in terms of the ith subject's mean response, which in this case is the ith subject's probability of heartburn relief. The coefficient β_1 is called a fixed effect, whereas b_{0i} is called a random effect. The random effect (b_{0i}) in this model is assumed to follow a normal distribution with mean 0 and variance $\sigma_{b_0}{}^2$. Subject-specific random effects are designed to account for the subject-to-subject variation, which may be due to unexplained genetic or environmental factors that are otherwise unaccounted for in the model. More generally, random effects are often used when levels of a variable are selected at random from a large population of possible levels.

For each subject:

logit of baseline risk = $(\beta_0 + b_{0i})$

b_{0i} = subject-specific intercept

With this model, each subject has his/her own baseline risk, the logit of which is the intercept plus the random effect $(\beta_0 + b_{0i})$. The sum $(\beta_0 + b_{0i})$ is typically called the subject-specific intercept. The amount of variation in the baseline risk is determined by the variance $({\sigma_{b_0}}^2)$ of b_{0i}.

No random effect for RX
$\Downarrow$
RX effect is same for each subject
i.e., $\exp(\beta_1)$

In addition to the intercept, we could have added another random effect allowing the treatment (RX) effect to also vary by subject (see Practice Exercises 4–9). By not adding this additional random effect, there is an assumption that the odds ratio for the effect of treatment is the same for each subject, $\exp(\beta_1)$.

Mixed logistic model (MLM)

```
                         Standard     Wald
Variable   Coefficient    Error     p-value

INTERCEPT  −0.2285        0.3583    0.5274
RX           0.3445       0.4425    0.4410
```

The output obtained from running the MLM on the heartburn data is presented on the left. This model was run using SAS's GLIMMIX procedure. (See the Computer Appendix for details and an example of program coding.)

Odds ratio and 95% CI:

$\widehat{\text{OR}} = \exp(0.3445) = 1.41$

$95\%\ \text{CI} = (0.593, 3.360)$

The odds ratio estimate for the effect of treatment for relieving heartburn is exp(0.3445) = 1.41. The 95% confidence interval is (0.593, 3.360).

Model comparison

Model	$\widehat{\text{OR}}$	$s_{\hat{\beta}}$
MLM	1.41	0.4425
GEE	1.35	0.3868
CLR	1.50	0.5271

The odds ratio estimate using this model is slightly larger than the estimate obtained (1.35) using the GEE approach, but somewhat smaller than the estimate obtained (1.50) using the conditional logistic regression approach. The standard error at 0.4425 is also larger than what was obtained in the GEE model (0.3868), but smaller than in the conditional logistic regression (0.5271).

Typical model for random Y:

- Fixed component (fixed effects)
- Random component (error)

The modeling of any response variable typically contains a fixed and random component. The random component, often called the error term, accounts for the variation in the response variables that the fixed predictors fail to explain.

Random effects model:

- Fixed component (fixed effects)
- Random components (random effects)
  1. Random effects: **b**
     Var(**b**) = **G**
  2. Residual variation: $\boldsymbol{\varepsilon}$
     Var($\boldsymbol{\varepsilon}$) = **R**

A model containing a random effect adds another layer to the random part of the model. With a random effects model, there are at least two random components in the model:

1. The first random component is the variation explained by the random effects. For the heartburn data set, the random effect is designed to account for random subject-to-subject variation (heterogeneity). The variance–covariance matrix of this random component (**b**) is called the **G** matrix.
2. The second random component is the residual error variation. This is the variation unexplained by the rest of the model (i.e., unexplained by fixed or random effects). For a given subject, this is the difference of the observed and expected response. The variance–covariance matrix of this random component is called the **R** matrix.

Random components layered:

$$Y_{ij} = \frac{1}{1+\exp\left[-\beta_0+\sum_{h=1}^{p}\beta_h X_{hij}+\underbrace{b_{oi}}_{\text{random effects}}\right]} + \underbrace{\boldsymbol{\varepsilon}_{ij}}_{\text{residual variation}}$$

For mixed logistic models, the layering of these random components is tricky. This layering can be illustrated by presenting the model (see left side) in terms of the random effect for the *i*th subject (b_{0i}) and the residual variation (ε_{ij}) for the *j*th response of the *i*th subject (Y_{ij}).

$$\varepsilon_{ij} = Y_{ij} - \mathrm{P}(Y_{ij} = 1|\mathbf{X}),$$

where

$$\mathrm{P}(Y_{ij} = 1|\mathbf{X}) = \mu$$
$$= \frac{1}{1 + \exp\left[-(\beta_0 + \sum_{h=1}^{p} \beta_h X_{hij} + b_{0i})\right]}$$

The residual variation (ε_{ij}) accounts for the difference between the observed value of Y and the mean response, $\mathrm{P}(Y = 1|\mathbf{X})$, for a given subject. The random effect (b_{0i}), on the other hand, allows for randomness in the modeling of the mean response [i.e., $\mathrm{P}(Y = 1|\mathbf{X})$], which is modeled using both fixed (βs) and random ($\mathrm{b_s}$) effects.

	GLM	**GEE**
Model	$Y_i = \mu_i + \varepsilon_{ij}$	$Y_{ij} = \mu_{ij} + \varepsilon_{ij}$
R	Independent	Correlated
G	—	—

For GLM and GEE models, the outcome Y is modeled as the sum of the mean and the residual variation [$Y = \mu + \varepsilon_{ij}$, where the mean ($\mu$) is fixed] determined by the subject's pattern of covariates. For GEE, the residual variation is modeled with a correlation structure, whereas for GLM, the residual variation (the **R** matrix) is modeled as independent. Neither GLM nor GEE models contain a **G** matrix, as they do not contain any random effects (**b**).

GLMM: $Y_{ij} = \underbrace{g^{-1}(\mathbf{X}, \beta, b_{0i})}_{\mu_{ij}} + \varepsilon_{ij}$

User specifies covariance structures for **R, G,** or both

In contrast, for GLMMs, the mean also contains a random component (b_{0i}). With GLMM, the user can specify a covariance structure for the **G** matrix (for the random effects), the **R** matrix (for the residual variation), or both. Even if the **G** and **R** matrices are modeled to contain zero correlations separately, the combination of both matrices in the model generally forms a correlated random structure for the response variable.

GEE: correlation structure specified

GLMM: covariance structure specified

Covariance structure contains parameters for both the variance and covariance

Another difference between a GEE model and a mixed model (e.g., MLM) is that a correlation structure is specified with a GEE model, whereas a covariance structure is specified with a mixed model. A covariance structure contains parameters for both the variance and covariance, whereas a correlation structure contains parameters for just the correlation. Thus, with a covariance structure, there are additional variance parameters and relationships among those parameters (e.g., variance heterogeneity) to consider (see Practice Exercises 7–9).

Covariance $\Rightarrow$ unique correlation

but correlation $\xcancel{\Rightarrow}$ unique covariance

If a covariance structure is specified, then the correlation structure can be ascertained. The reverse is not true, however, since a correlation matrix does not in itself determine a unique covariance matrix.

For ith subject:

- **R** matrix dimensions depend on number of observations for subject i (n_i)
- **G** matrix dimensions depend on number of random effects (q)

Heartburn data:

R $= 2 \times 2$ matrix

G $= 1 \times 1$ matrix (only one random effect)

For a given subject, the dimensions of the **R** matrix depend on how many observations (n_i) the subject contributes, whereas the dimensions of the **G** matrix depend on the number of random effects (e.g., q) included in the model. For the heartburn data example, in which there are two observations per subject ($n_i = 2$), the **R** matrix is a 2×2 matrix modeled with zero correlation. The dimensions of **G** are 1×1 ($q = 1$) since there is only one random effect (b_{0i}), so there is no covariance structure to consider for **G** in this model. Nevertheless, the combination of the **G** and **R** matrix in this model provides a way to account for correlation between responses from the same subject.

CLR vs. MLM: subject-specific effects

CLR: logit $\mu_i = \beta_0 + \beta_1\text{RX} + \gamma_i$, where γ_i is a *fixed* effect

MLM: logit $\mu_i = \beta_0 + \beta_1\text{RX} + b_{0i}$, where b_{0i} is a *random* effect

We can compare the modeling of subject-specific fixed effects with subject-specific random effects by examining the conditional logistic model (CLR) and the mixed logistic model (MLM) in terms of the ith subject's response. Using the heartburn data, these models can be expressed as shown on the left.

Fixed effect γ_i: impacts modeling of μ

Random effect b_{0i}: used to characterize the variance

The fixed effect, γ_i, impacts the modeling of the mean response. The random effect, b_{0i}, is a random variable with an expected value of zero. Therefore, b_{0i} does not directly contribute to the modeling of the mean response; rather, it is used to characterize the variance of the mean response.

GEE vs. MLM

GEE model: $\text{logit}\ \mu = \beta_0 + \beta_1\text{RX} \quad \bigcirc$

No subject-specific random effects (b_{0i})

Within-subject correlation specified in **R** matrix

MLM model: $\text{logit}\ \mu_i = \beta_0 + \beta_1\text{RX} + \boxed{b_{0i}}$

Subject-specific random effects

A GEE model can also be expressed in terms of the ith subject's mean response (μ_i), as shown at left using the heartburn example. The GEE model contrasts with the MLM, and the conditional logistic regression, since the GEE model does not contain subject-specific effects (fixed or random). With the GEE approach, the within-subject correlation is handled by the specification of a correlation structure for the **R** matrix. However, the mean response is not directly modeled as a function of the individual subjects.

Marginal model ⇒ $E(Y|\mathbf{X})$ not conditioned on cluster-specific information

(e.g., *not* allowed as X

- Earlier values of Y
- Subject-specific effects)

A GEE model represents a type of model called a marginal model. With a marginal model, the mean response $E(Y|\mathbf{X})$ is not directly conditioned on any variables containing information on the within-cluster correlation. For example, the predictors ($\mathbf{X}$) in a marginal model cannot be earlier values of the response from the same subject or subject-specific effects.

Marginal models (examples):

GEE
ALR
SLR

Other examples of marginal models include the ALR model, described earlier in the chapter, and the standard logistic regression with one observation for each subject. In fact, any model using data in which there is one observation per subject is a marginal model because in that situation, there is no information available about within-subject correlation.

Heartburn Relief Study

β_1 = parameter of interest

BUT

interpretation of exp (β_1) depends on type of model

Returning to the Heartburn Relief Study example, the parameter of interest is the coefficient of the RX variable, β_1, not the subject-specific effect, b_{0i}. The research question for this study is whether the active treatment provides greater relief for heartburn than the standard treatment. The interpretation of the odds ratio exp(β_1) depends, in part, on the type of model that is run.

Heartburn Relief Study:

GEE: marginal model

$\exp(\hat{\beta}_1)$ is *population* $\widehat{\text{OR}}$

MLM:

$\exp(\hat{\beta}_1)$ is $\widehat{\text{OR}}$ for an *individual*

The odds ratio for a marginal model is the ratio of the odds of heartburn for RX = 1 vs. RX = 0 among the *underlying population*. In other words, the OR is a population average. The odds ratio for a model with a subject-specific effect, as in the mixed logistic model, is the ratio of the odds of heartburn for RX = 1 vs. RX = 0 for an *individual*.

What is an individual OR?

Each subject has separate probabilities

$\text{P}(X = 1|\text{RX} = 1)$

$\text{P}(X = 1|\text{RX} = 0)$

$\Downarrow$

OR compares RX = 1 vs. RX = 0 for an individual

What is meant by an odds ratio for an individual? We can conceptualize each subject as having a probability of heartburn relief given the active treatment and having a separate probability of heartburn relief given the standard treatment. These probabilities depend on the fixed treatment effect as well as the subject-specific random effect. With this conceptualization, the odds ratio that compares the active vs. standard treatment represents a parameter that characterizes an individual rather than a population (see Practice Exercises 10–15). The mixed logistic model supplies a structure that gives the investigator the ability to estimate an odds ratio for an individual, while simultaneously accounting for within-subject and between-subject variation.

Goal		OR
Population inferences	⇒	marginal
Individual inferences	⇒	individual

The choice of whether a population averaged or individual level odds ratio is preferable depends, in part, on the goal of the study. If the goal is to make inferences about a population, then a marginal effect is preferred. If the goal is to make inferences on the individual, then an individual level effect is preferred.

Parameter estimation for MLM in SAS:

GLIMMIX

- Penalized quasi-likelihood equations
- User specifies **G** and **R**

NLMIXED

- Maximized approximation to likelihood integrated over random effects
- User does not specify **G** and **R**
- User specifies variance components of **G** matrix and assumes an independent **R** matrix (i.e., $\mathbf{R} = \sigma^2 I$)

There are various methods that can be used for parameter estimation with mixed logistic models. The parameter estimates, obtained for the Heartburn Relief data from the SAS procedure GLIMMIX use an approach termed penalized quasi-likelihood equations (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993). Alternatively, the SAS procedure NLMIXED can also be used to run a mixed logistic model. NLMIXED fits nonlinear mixed models by maximizing an approximation to the likelihood integrated over the random effects. Unlike GLIMMIX, NLMIXED does not allow the user to specify a correlation structure for the **G** and **R** matrices (SAS Institute, 2000).

Instead, NLMIXED allows the user to specify the individual variance components within the **G** matrix, but assumes that the **R** matrix has an independent covariance structure (i.e. 0s on the off-diagonals of the **R** matrix).

Mixed models are flexible:

- Layer random components
- Handle nested clusters
- Control for subject effects

Mixed models offer great flexibility by allowing the investigator to layer random components, model clusters nested within clusters (i.e., perform hierarchical modeling), and control for subject-specific effects. The use of mixed *linear* models is widespread in a variety of disciplines because of this flexibility.

Performance of mixed logistic models not fully evaluated

Despite the appeal of mixed *logistic* models, their performance, particularly in terms of numerical accuracy, has not yet been adequately evaluated. In contrast, the GEE approach has been thoroughly investigated, and this is the reason for our emphasis on that approach in the earlier chapters on correlated data (Chaps. 14 and 15).

V. SUMMARY

✓Chapter 16. Other Approaches for Analysis of Correlated Data

The presentation is now complete. Several alternate approaches for the analysis of correlated data were examined and compared to the GEE approach. The approaches discussed included alternating logistic regressions, conditional logistic regression, and the generalized linear mixed (logistic) model.

The choice of which approach to implement for the primary analysis can be difficult and should be determined, in part, by the research hypothesis. It may be of interest to use several different approaches for comparison. If the results are different, it can be informative to investigate why they are different. If they are similar, it may be reassuring to know the results are robust to different methods of analysis.

We suggest that you review the material covered here by reading the detailed outline that follows.

Computer Appendix

A Computer Appendix is presented in the following section. This appendix provides details on performing the analyses discussed in the various chapters using SAS, SPSS, and Stata statistical software.

Detailed Outline

I. Overview (page 570)

A. Other approaches for analysis of correlated data:

i. Alternating logistic regressions (ALR) algorithm

ii. Conditional logistic regression

iii. Generalized linear mixed model (GLMM)

II. Alternating logistic regressions algorithm (pages 571–575)

A. Similar to GEE except that

i. Associations between pairs of responses are modeled with odds ratios instead of correlations:

$$\mathrm{OR}_{ijk} = \frac{\mathrm{P}(Y_{ij}=1, Y_{ik}=1)\mathrm{P}(Y_{ij}=0, Y_{ik}=0)}{\mathrm{P}(Y_{ij}=1, Y_{ik}=0)\mathrm{P}(Y_{ij}=0, Y_{ik}=1)}.$$

ii. Associations between responses may also be of interest, and not considered nuisance parameters.

III. Conditional logistic regression (pages 575–579)

A. May be applied in a design where each subject can be viewed as a stratum (e.g., has an exposed and an unexposed observation).

B. Subject-specific fixed effects are included in the model through the use of dummy variables [Example: Heartburn Relief Study ($n = 40$)]:

$$\text{logit } \mathrm{P}(\mathbf{X}) = \beta_0 + \beta_1\mathrm{RX} + \sum_{i=1}^{39} \gamma_i V_i,$$

where $V_i = 1$ for subject i and $V_i = 0$ otherwise.

C. In the output, there are no parameter estimates for the intercept or the dummy variables representing the matched factor, as these parameters cancel out in the conditional likelihood.

D. An important distinction between CLR and GEE is that a matched analysis (CLR) relies on the within-subject variability in the estimation of the parameters, whereas a correlated analysis (GEE) relies on both the within-subject variability and the between-subject variability.

IV. The generalized linear mixed model approach (pages 579–587)

A. A generalization of the linear mixed model.

B. As with the GEE approach, the user can specify the logit link function and apply a variety of covariance structures to the model.

C. As with the conditional logistic regression approach, subject-specific effects are included in the model:

   logit $\mu_i = \mathrm{P}(D = 1|\mathrm{RX}) = \beta_0 + \beta_1\mathrm{RX}_i + b_{0i}$,

   where b_i is a random variable from a normal distribution with mean = 0 and variance = ${\sigma_{b_0}}^2$.

D. Comparing the conditional logistic model and the mixed logistic model:

   i. The conditional logistic model:

      logit $\mu_i = \beta_0 + \beta_1\mathrm{RX} + \gamma_i$, where γ_i is a fixed effect

   ii. The mixed logistic model:

      logit $\mu_i = \beta_0 + \beta_1\mathrm{RX} + \mathrm{b}_{0i}$,

      where b_{0i} is a random effect.

E. Interpretation of the odds ratio:

   i. Marginal model: population average OR

   ii. Subject-specific effects model: individual OR.

V. **Summary** (page 588)

Practice Exercises

The Practice Exercises presented here are primarily designed to elaborate and expand on several concepts that were briefly introduced in this chapter.

Exercises 1–3 relate to calculating odds ratios and their corresponding correlations. Consider the following 2 × 2 table for two dichotomous responses (Y_j and Y_k). The cell counts are represented by A, B, C, and D. The margins are represented by M_1, M_0, N_1, and N_0 and the total counts are represented by T.

	$Y_k = 1$	$Y_k = 0$	Total
$Y_j = 1$	A	B	$M_1 = A + B$
$Y_j = 0$	C	D	$M_0 = C + D$
Total	$N_1 = A + C$	$N_0 = B + D$	$T = A + B + C + D$

The formulas for calculating the correlation and odds ratio between Y_j and Y_k in this setting are given as follows:

$$\text{Corr}(Y_j, Y_k) = \frac{AT - M_1N_1}{\sqrt{M_1M_0N_1N_0}}, \quad \text{OR} = \frac{AD}{BC}.$$

1. Calculate and compare the respective odds ratios and correlations between Y_j and Y_k for the data summarized in Tables 1 and 2 to show that the same odds ratio can correspond to different correlations.

Table 1

	$Y_k = 1$	$Y_k = 0$
$Y_j = 1$	3	1
$Y_j = 0$	1	3

Table 2

	$Y_k = 1$	$Y_k = 0$
$Y_j = 1$	9	1
$Y_j = 0$	1	1

2. Show that if *both* the B and C cells are equal to 0, then the correlation between Y_j and Y_k is equal to 1 (assuming A and D are nonzero). What is the corresponding odds ratio if the B and C cells are equal to 0? Did both the B and C cells have to equal 0 to obtain this corresponding odds ratio? Also show that if *both* the A and D cells are equal to zero, then the correlation is equal to −1. What is that corresponding odds ratio?
3. Show that if $AD = BC$, then the correlation between Y_j and Y_k is 0 and the odds ratio is 1 (assuming nonzero cell counts).

Exercises 4–6 refer to a model constructed using the data from the Heartburn Relief Study. The dichotomous outcome is relief from heartburn (coded 1 = yes, 0 = no). The only predictor variable is RX (coded 1 = active treatment, 0 = standard treatment). This model contains *two* subject-specific effects: one for the intercept (b_{0i}) and the other (b_{1i}) for the coefficient RX. The model is stated in terms of the ith subject's mean response:

$$\text{logit } \mu_i = \text{P}(D = 1|\text{RX}) = \beta_0 + \beta_1\text{RX}_i + b_{0i} + b_{1i}\text{RX}_i,$$

where b_{0i} follows a normal distribution with mean 0 and variance $\sigma_{b_0}{}^2$, b_{1i} follows a normal distribution with mean 0 and variance $\sigma_{b_1}{}^2$ and where the covariance matrix of b_{0i} and b_{1i} is a 2 × 2 matrix, **G**.

It may be helpful to restate the model by rearranging the parameters such that the intercept parameters (fixed and random effects) and the slope parameters (fixed and random effects) are grouped together:

$$\text{logit } \mu_i = \text{P}(D = 1|\text{RX}) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\text{RX}_i$$

4. Use the model to obtain the odds ratio for RX = 1 vs. RX = 0 for subject i.
5. Use the model to obtain the baseline *risk* for subject i (i.e., risk when RX = 0).
6. Use the model to obtain the odds ratio (RX = 1 vs. RX = 0) averaged over all subjects.

Below are three examples of commonly used covariance structures represented by 3 × 3 matrices. The elements are written in terms of the variance (σ^2), standard deviation (σ), and correlation (ρ). The covariance structures are presented in this form in order to contrast their structures with the correlation structures presented in Chap. 14. A covariance structure not only contains correlation parameters but variance parameters as well.

Variance components — Compound symmetric — Unstructured

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} \begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2\rho & \sigma^2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \sigma_1\sigma_3\rho_{13} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 & \sigma_2\sigma_3\rho_{23} \\ \sigma_1\sigma_3\rho_{13} & \sigma_2\sigma_3\rho_{23} & \sigma_3^2 \end{bmatrix}$$

The compound symmetric covariance structure has the additional constraint that $\rho \geq 0$,

7. Which of the above covariance structures allow for variance heterogeneity within a cluster?
8. Which of the presented covariance structures allow for both variance heterogeneity and correlation within a cluster.
9. Consider a study in which there are five responses per subject. If a model contains two subject-specific random effects (for the intercept and slope), then for subject i, what are the dimensions of the **G** matrix and of the **R** matrix?

The next set of exercises is designed to illustrate how an individual level odds ratio can differ from a population averaged (marginal) odds ratio. Consider a fictitious data set in which there are only 2 subjects, contributing 200 observations apiece. For each subject, 100 of the observations are exposed ($E = 1$) and 100 are unexposed ($E = 0$), yielding 400 total observations. The outcome is dichotomous ($D = 1$ and $D = 0$). The data are summarized using three 2 × 2 tables. The tables for Subject 1 and Subject 2 summarize the data for each subject; the third table pools the data from both subjects.

Subject 1		
	$E=1$	$E=0$
$D=1$	50	25
$D=0$	50	75
Total	100	100

Subject 2		
	$E = 1$	$E = 0$
$D=1$	25	10
$D=0$	75	90
Total	100	100

Pooled subjects		
	$E = 1$	$E = 0$
$D=1$	75	35
$D=0$	125	165
Total	200	200

10. Calculate the odds ratio for Subject 1 and Subject 2 separately. Calculate the odds ratio after pooling the data for both subjects. How do the odds ratios compare?

    Note: The subject-specific odds ratio as calculated here is a conceptualization of a subject-specific effect, while the pooled odds ratio is a conceptualization of a population-averaged effect.
11. Compare the baseline risk (where $E = 0$) of Subject 1 and Subject 2. Is there a difference (i.e., heterogeneity) in the baseline risk between subjects? Note that for a model containing subject-specific random effects, the variance of the random intercept is a measure of baseline risk heterogeneity.
12. Do Subject 1 and Subject 2 have a different distribution of exposure? This is a criterion for evaluating whether there is confounding by subject.
13. Suppose an odds ratio is estimated using data in which there are many subjects, each with one observation per subject. Is the odds ratio estimating an individual level odds ratio or a population averaged (marginal) odds ratio?

For Exercise 14 and Exercise 15, consider a similar scenario as was presented above for Subject 1 and Subject 2. However, this time the risk ratio rather than the odds ratio is the measure of interest. The data for Subject 2 have been altered slightly in order to make the risk ratio the same for each subject allowing comparability to the previous example.

Subject 1		
	$E=1$	$E=0$
$D=1$	50	25
$D=0$	50	75
Total	100	100

Subject 2		
	$E=1$	$E=0$
$D=1$	20	10
$D=0$	80	90
Total	100	100

Pooled subjects		
	$E=1$	$E=0$
$D=1$	70	35
$D=0$	130	165
Total	200	200

14. Compare the baseline risk (where $E = 0$) of Subject 1 and Subject 2. Is there a difference (i.e., heterogeneity) in the baseline risk between subjects?
15. Calculate the risk ratio for Subject 1 and Subject 2 separately. Calculate the risk ratio after pooling the data for both subjects. How do the risk ratios compare?

Test

True or false (Circle T or F)

T F 1. A model with subject-specific random effects is an example of a marginal model.

T F 2. A conditional logistic regression cannot be used to obtain parameter estimates for a predictor variable that does not vary its values within the matched cluster.

T F 3. The alternating logistic regressions approach models relationships between pairs of responses from the same cluster with odds ratio parameters rather than with correlation parameters as with GEE.

T F 4. A mixed logistic model is a generalization of the generalized linear mixed model in which a link function can be specified for the modeling of the mean.

T F 5. For a GEE model, the user specifies a correlation structure for the response variable, whereas for a GLMM, the user specifies a covariance structure.

Questions 6–10 refer to models run on the data from the Heartburn Relief Study. The following printout summarizes the computer output for two mixed logistic models. The models include a subject-specific random effect for the intercept. The dichotomous outcome is relief from heartburn (coded 1 = yes, 0 = no). The exposure of interest is RX (coded 1 = active treatment, 0 = standard treatment). The variable SEQUENCE is coded 1 for subjects in which the active treatment was administered first and 0 for subjects in which the standard treatment was administered first. The product term RX*SEQ (RX times SEQUENCE) is included to assess interaction between RX and SEQUENCE. Only the estimates of the fixed effects are displayed in the output.

Model 1

```
Variable         Estimate         Std Err

INTERCEPT        −0.6884          0.5187
RX                0.4707          0.6608
SEQUENCE          0.9092          0.7238
RX*SEQ           −0.2371          0.9038
```

Model 2

```
Variable       Coefficient        Std Err

INTERCEPT      −0.6321            0.4530
RX              0.3553            0.4565
SEQUENCE        0.7961            0.564
```

6. State the logit form of Model 1 in terms of the mean response of the ith subject.
7. Use Model 1 to estimate the odds ratios for RX (active vs. standard treatment).
8. Use Model 2 to estimate the odds ratio and 95% confidence intervals for RX.
9. How does the interpretation of the odds ratio for RX using Model 2 compare to the interpretation of the odds ratio for RX using a GEE model with the same covariates (see Model 2 in the Test questions of Chap. 15)?
10. Explain why the parameter for SEQUENCE cannot be estimated in a conditional logistic regression using subject as the matching factor.

Answers to Practice Exercises

1. The odds ratio for the data in Table 1 and Table 2 is 9. The correlation for the data in Table 1 is $\frac{[(3)(8)-(4)(4)]}{\sqrt{(4)(4)(4)(4)}} = 0.5$, and for the data in Table 2, it is $\frac{[(9)(12)-(10)(10)]}{\sqrt{(10)(2)(10)(2)}} = 0.4$ So, the odds ratios are the same but the correlations are different.
2. If $B = C = 0$, then $M_1 = N_1 = A$ and $M_0 = N_0 = D$ and $T = A + D$.

   $$\text{corr} = \frac{AT - M_1N_1}{\sqrt{M_1M_0N_1N_0}} = \frac{A(A+D) - AD}{\sqrt{(AD)(AD)}} = 1.$$

   The corresponding odds ratio is infinity. Even if just one of the cells (B or C) were zero, the odds ratio would still be infinity, but the corresponding correlation would be less than 1.
   If $A = D = 0$, then $M_1 = N_0 = B$ and $M_0 = N_1 = C$ and $T = B + C$.

   $$\text{corr} = \frac{AT - M_1N_1}{\sqrt{M_1M_0N_1N_0}} = \frac{0(B+C) - BC}{\sqrt{(BC)(BC)}} = -1.$$

   The corresponding odds ratio is zero.
3. If $AD = BC$, then $D = (BC)/A$ and $T = A + B + C + (BC)/A$.

   $$\text{corr} = \frac{AT - M_1N_1}{\sqrt{M_1M_0N_1N_0}}$$
   $$= \frac{A[A + B + C + (BC/A)] - [(A+B)(A+C)]}{\sqrt{M_1M_0N_1N_0}} = 0$$

   $\text{OR} = \frac{AD}{BC} = \frac{AD}{AD} = 1$ (indicating no association between Y_j and Y_k).
4. $\exp(\beta_1 + b_{1i})$
5. $\dfrac{1}{1 + \exp[-(\beta_0 + b_{0i})]}$
6. $\exp(\beta_1)$ since b_{1i} is a random variable with a mean of 0 (compare to Exercise 4).
7. The variance components and unstructured covariance structures allow for variance heterogeneity.
8. The unstructured covariance structure.
9. The dimensions of the **G** matrix are 2 × 2 and the dimensions of the **R** matrix are 5 × 5 for subject i.
10. The odds ratio is 3.0 for both Subject 1 and Subject 2 separately, whereas the odds ratio is 2.83 after pooling the data. The pooled odds ratio is smaller.
11. The baseline risk for Subject 1 is 0.25, whereas the baseline risk for Subject 2 is 0.10. There is a difference in the baseline risk (although there are only two subjects). *Note*. In general, assuming there is heterogeneity

in the subject-specific baseline risk, the population averaging for a marginal odds ratio attenuates (i.e., weakens) the effect of the individual level odds ratio.

12. Subject 1 and Subject 2 have the same distribution of exposure: 100 exposed out of 200 observations. *Note*. In a case-control setting we would consider that there is a different distribution of exposure where $D = 0$.
13. With one observation per subject, an odds ratio is estimating a population-averaged (marginal) odds ratio since in that setting observations must be pooled over subjects.
14. The baseline risk for Subject 1 is 0.25, whereas the baseline risk for Subject 2 is 0.10, indicating that there is heterogeneity of the baseline risk between subjects.
15. The risk ratio is 2.0 for both Subject 1 and Subject 2 separately and the risk ratio is also 2.0 for the pooled data. In contrast to the odds ratio, if the distribution of exposure is the same across subjects, then pooling the data does not attenuate the risk ratio in the presence of heterogeneity of the baseline risk.

Appendix: Computer Programs for Logistic Regression

In this appendix, we provide examples of computer programs to carry out unconditional logistic regression, conditional logistic regression, polytomous logistic regression, ordinal logistic regression, and GEE logistic regression. This appendix does not give an exhaustive survey of all computer packages currently available, but rather is intended to describe the similarities and differences among a sample of the most widely used packages. The software packages that we consider are SAS version 9.2, SPSS version 16.0, and Stata version 10.0. A detailed description of these packages is beyond the scope of this appendix. Readers are referred to the built-in Help functions for each program for further information.

The computer syntax and output presented in this appendix are obtained from running models on five datasets. We provide each of these datasets on an accompanying disk in four forms: (1) as text datasets (with a **.dat** extension), (2) as SAS version 9 datasets (with a **.sas7bdat** extension), (3) as SPSS datasets (with a **.sav** extension), and (4) as Stata datasets (with a **.dta** extension). Each of the four datasets is described below. We suggest making backup copies of the datasets prior to use to avoid accidentally overwriting the originals.

DATASETS

1. Evans County dataset (**evans.dat**)

The **evans.dat** dataset is used to demonstrate a standard logistic regression (unconditional). The Evans County dataset is discussed in Chap. 2. The data are from a cohort study in which 609 white males were followed for 7 years, with coronary heart disease as the outcome of interest. The variables are defined as follows:

ID – The subject identifier. Each observation has a unique identifier since there is one observation per subject.

CHD – A dichotomous outcome variable indicating the presence (coded 1) or absence (coded 0) of coronary heart disease.

CAT – A dichotomous predictor variable indicating high (coded 1) or normal (coded 0) catecholamine level.
AGE – A continuous variable for age (in years).
CHL – A continuous variable for cholesterol.
SMK – A dichotomous predictor variable indicating whether the subject ever smoked (coded 1) or never smoked (coded 0).
ECG – A dichotomous predictor variable indicating the presence (coded 1) or absence (coded 0) of electrocardiogram abnormality.
DBP – A continuous variable for diastolic blood pressure.
SBP – A continuous variable for systolic blood pressure.
HPT – A dichotomous predictor variable indicating the presence (coded 1) or absence (coded 0) of high blood pressure. HPT is coded 1 if the systolic blood pressure is greater than or equal to 160 or the diastolic blood pressure is greater than or equal to 95.
CH and CC – Product terms of CAT × HPT and CAT × CHL, respectively.

2. MI dataset (**mi.dat**)

This dataset is used to demonstrate conditional logistic regression. The MI dataset is discussed in Chap. 11. The study is a case-control study that involves 117 subjects in 39 matched strata. Each stratum contains three subjects, one of whom is a case diagnosed with myocardial infarction while the other two are matched controls. The variables are defined as follows:

MATCH – A variable indicating the subject's matched stratum. Each stratum contains one case and two controls and is matched on age, race, sex, and hospital status.
PERSON – The subject identifier. Each observation has a unique identifier since there is one observation per subject.
MI – A dichotomous outcome variable indicating the presence (coded 1) or absence (coded 0) of myocardial infarction.
SMK – A dichotomous variable indicating whether the subject is (coded 1) or is not (coded 0) a current smoker.
SBP – A continuous variable for systolic blood pressure.
ECG – A dichotomous predictor variable indicating the presence (coded 1) or absence (coded 0) of electrocardiogram abnormality.

3. Cancer dataset (**cancer.dat**)

This dataset is used to demonstrate polytomous and ordinal logistic regression. The cancer dataset, discussed in Chaps. 12 and 13, is part of a study of cancer survival (Hill et al., 1995). The study involves 288 women who had been diagnosed with endometrial cancer. The variables are defined as follows:

ID – The subject identifier. Each observation has a unique identifier since there is one observation per subject.
GRADE – A three-level ordinal outcome variable indicating tumor grade. The grades are well differentiated (coded 0), moderately differentiated (coded 1), and poorly differentiated (coded 2).
RACE – A dichotomous variable indicating whether the race of the subject is black (coded 1) or white (coded 0).

ESTROGEN – A dichotomous variable indicating whether the subject ever (coded 1) or never (coded 0) used estrogen.

SUBTYPE – A three-category polytomous outcome indicating whether the subject's histological subtype is Adenocarcinoma (coded 0), Adenosquamous (coded 1), or Other (coded 2).

AGE – A dichotomous variable indicating whether the subject is within the age group 50–64 (coded 0) or within the age group 65–79 (coded 1). All 286 subjects are within one of these age groups.

SMK – A dichotomous variable indicating whether the subject is (coded 1) or is not (coded 0) a current smoker.

4. Infant dataset (**infant.dat**)

This is the dataset that is used to demonstrate GEE modeling. The infant dataset, discussed in Chaps. 14 and 15, is part of a health intervention study in Brazil (Cannon et al., 2001). The study involves 168 infants, each of whom has at least five and up to nine monthly measurements, yielding 1,458 observations in all. There are complete data on all covariates for 136 of the infants. The outcome of interest is derived from a weight-for-height standardized score based on the weight-for-height distribution of a standard population. The outcome is correlated since there are multiple measurements for each infant. The variables are defined as follows:

IDNO – The subject (infant) identifier. Each subject has up to nine observations. This is the variable that defines the cluster used for the correlated analysis.

MONTH – A variable taking the values 1 through 9 that indicates the order of an infant's monthly measurements. This is the variable that distinguishes observations within a cluster.

OUTCOME – Dichotomous outcome of interest derived from a weight-for-height standardized z-score. The outcome is coded 1 if the infant's z-score for a particular monthly measurement is less than negative one and coded 0 otherwise.

BIRTHWGT – A continuous variable that indicates the infant's birth weight in grams. This is a time-independent variable, as the infant's birth weight does not change over time. The value of the variable is missing for 32 infants.

GENDER – A dichotomous variable indicating whether the infant is male (coded 1) or female (coded 2).

DIARRHEA – A dichotomous time-dependent variable indicating whether the infant did (coded 1) or did not (coded 0) have symptoms of diarrhea that month.

5. Knee Fracture dataset (**kneefr.dat**)

This dataset is used to demonstrate how to generate classification tables and receiver operating characteristic (ROC) curves using logistic regression. The knee fracture dataset discussed in Chap. 10 contains information on 348 patients of which 45 actually had a knee fracture (Tigges et al., 1999). The goal of the study is to evaluate whether a patient's pattern of covariates can be used as a screening test before performing the X-ray. Since 1.3 million people visit North American emergency departments annually complaining of blunt knee trauma, the total cost associated with even a relatively inexpensive test such as a knee radiograph may be substantial.

The variables are defined as follows:

FRACTURE – A dichotomous variable coded 1 for a knee fracture, 0 for no knee fracture (obtained from X-ray)

FLEX – A dichotomous variable for the ability to flex the knee, coded 0 = yes, 1 = no

WEIGHT – A dichotomous variable for the ability to put weight on the knee, coded 0 = yes, 1 = no

AGECAT – A dichotomous variable for patient's age, coded 0 if age <55, 1 if age ≥ 55

HEAD – A dichotomous variable for injury to the knee head, coded 0 = no, 1 = yes

PETELLAR – A dichotomous variable for injury to the patellar, coded 0 = no, 1 = yes

We first illustrate how to perform analyses of these datasets using SAS, followed by SPSS, and finally Stata. Not all of the output produced from each procedure will be presented, as some of the output is extraneous to our discussion.

SAS

Analyses are carried out in SAS by using the appropriate SAS procedure on a SAS dataset. Each SAS procedure begins with the word PROC. The following SAS procedures are used to perform the analyses in this appendix.

1) PROC LOGISTIC – This procedure can be used to run logistic regression (unconditional and conditional), general polytomous logistic regression, and ordinal logistic regression using the proportional odds model.
2) PROC GENMOD – This procedure can be used to run generalized linear models (GLM – including unconditional logistic regression and ordinal logistic regression) and GEE models.
3) PROC GLIMMIX – This procedure can be used to run generalized linear mixed models (GLMMs).

The capabilities of these procedures are not limited to performing the analyses listed above. However, our goal is to demonstrate only the types of modeling presented in this text.

Unconditional Logistic Regression

A. PROC LOGISTIC

The first illustration presented is an unconditional logistic regression with PROC LOGISTIC using the Evans County data. The dichotomous outcome variable is CHD and the predictor variables are: CAT, AGE, CHL, ECG, SMK, and HPT. Two interaction terms, CH and CC, are also included. CH is the product: CAT × HPT, while CC is the product: CAT × CHL. The variables representing the interaction terms have already been included in the datasets.

The model is stated as follows:

$$\begin{aligned}\text{logit } P(\text{CHD} = 1|\mathbf{X}) = \beta_0 &+ \beta_1\text{CAT} + \beta_2\text{AGE} + \beta_3\text{CHL} + \beta_4\text{ECG} + \beta_5\text{SMK} \\ &+ \beta_6\text{HPT} + \beta_7\text{CH} + \beta_8\text{CC}\end{aligned}$$

For this example, we shall use the SAS permanent dataset **evans.sas7bdat**. A LIBNAME statement is needed to indicate the path to the location of the SAS dataset. In our examples, we assume the file is located on the C drive. The LIBNAME statement includes a reference name as well as the path. We call the reference name REF. The code is as follows:

```
LIBNAME REF ‘C:\’;
```

The user is free to define his/her own reference name. The path to the location of the file is given between the single quotation marks. The general form of the code is

```
LIBNAME  Your reference name  ‘Your path to file location’;
```

All of the SAS programming will be written in capital letters for readability. However, SAS is *not* case sensitive. If a program is written with lower case letters, SAS reads them as upper case. The number of spaces between words (if more than one) has no effect on the program. Each SAS programming statement ends with a semicolon.

The code to run a standard logistic regression with PROC LOGISTIC is as follows:

```
PROC LOGISTIC DATA = REF.EVANS DESCENDING;
MODEL CHD = CAT AGE CHL ECG SMK HPT CH CC / COVB;
RUN;
```

With the LIBNAME statement, SAS recognizes a two-level file name: the reference name and the file name without an extension. For our example, the SAS file name is REF.EVANS. Alternatively, a temporary SAS dataset could be created and used. However, a temporary SAS dataset has to be recreated in every SAS session as it is deleted from memory when the user exits SAS. The following code creates a temporary SAS dataset called EVANS from the permanent SAS dataset REF.EVANS.

```
DATA EVANS;
SET REF.EVANS;
RUN;
```

The DESCENDING option in the PROC LOGISTIC statement instructs SAS that the outcome event of interest is CHD = 1 rather than the default, CHD = 0. In other words, we are interested in modeling the P(CHD = 1) rather than the P(CHD = 0). Check the Response Profile in the output to see that CHD = 1 is listed before CHD = 0. In general, if the output produces results that are the opposite of what you would expect, chances are that there is an error in coding, such as incorrectly omitting (or incorrectly adding) the DESCENDING option.

Options requested in the MODEL statement are preceded by a forward slash. The COVB option requests the variance–covariance matrix for the parameter estimates.

The output produced by PROC LOGISTIC follows:

```
                          The LOGISTIC Procedure

                            Model Information

        Data Set                              REF.EVANS
        Response Variable                        chd
        Number of Response Levels                 2
        Number of Observations                   609
        Link Function                           Logit
        Optimization Technique             Fisher's scoring

                             Response Profile

                     Ordered
                      Value         CHD        Count
                     1               1           71
                     2               0          538

                           Model Fit Statistics

                            Intercept        Intercept and
          Criterion           Only            Covariates
          AIC               440.558            365.230
          SC                444.970            404.936
          −2 Log L          438.558            347.230

                 Analysis of Maximum Likelihood Estimates

                            Standard
Parameter      DF          Estimate       Error       Chi-Square      Pr > ChiSq
Intercept       1          −4.0497       1.2550        10.4125         0.0013
CAT             1         −12.6894       3.1047        16.7055         <.0001
AGE             1           0.0350       0.0161         4.6936         0.0303
CHL             1          −0.00545      0.00418        1.7000         0.1923
ECG             1           0.3671       0.3278         1.2543         0.2627
SMK             1           0.7732       0.3273         5.5821         0.0181
HPT             1           1.0466       0.3316         9.9605         0.0016
CH              1          −2.3318       0.7427         9.8579         0.0017
CC              1           0.0692       0.0144        23.2020         <.0001

                          Odds Ratio Estimates

                        Point              95% Wald
        Effect        Estimate         Confidence Limits
        CAT             <0.001         <0.001         0.001
        AGE              1.036          1.003         1.069
        CHL              0.995          0.986         1.003
        ECG              1.444          0.759         2.745
        SMK              2.167          1.141         4.115
        HPT              2.848          1.487         5.456
        CH               0.097          0.023         0.416
        CC               1.072          1.042         1.102
```

```
                          Estimated Covariance Matrix

 Variable    Intercept       cat          age          chl          ecg

Intercept     1.575061   −0.66288     −0.01361     −0.00341     −0.04312
CAT          −0.66288     9.638853    −0.00207      0.003591     0.02384
AGE          −0.01361    −0.00207      0.00026     −3.66E-6      0.00014
CHL          −0.00341     0.003591    −3.66E-6      0.000018     0.000042
ECG          −0.04312     0.02384      0.00014      0.000042     0.107455
SMK          −0.1193     −0.02562      0.000588     0.000028     0.007098
HPT           0.001294    0.001428    −0.00003     −0.00025     −0.01353
CH            0.054804   −0.00486     −0.00104      0.000258    −0.00156
CC            0.003443   −0.04369      2.564E-6    −0.00002     −0.00033

    Variable        smk          hpt          ch           cc

    Intercept    −0.1193       0.001294     0.054804     0.003443
    CAT          −0.02562      0.001428    −0.00486     −0.04369
    AGE           0.000588    −0.00003     −0.00104      2.564E-6
    CHL           0.000028    −0.00025      0.000258    −0.00002
    ECG           0.007098    −0.01353     −0.00156     −0.00033
    SMK           0.107104    −0.00039      0.002678     0.000096
    HPT          −0.00039      0.109982    −0.108        0.000284
    CH            0.002678    −0.108        0.551555    −0.00161
    CC            0.000096     0.000284    −0.00161      0.000206
```

The negative 2 log likelihood statistic (i.e., −2 Log L) for the model, 347.230, is presented in the table titled “Model Fit Statistics.” A likelihood ratio test statistic to assess the significance of the two interaction terms can be obtained by running a no-interaction model and subtracting the negative 2 log likelihood statistic for the current model from that of the no-interaction model.

The parameter estimates are given in the table titled “Analysis of Maximum Likelihood Estimates.” The point estimates of the odds ratios, given in the table titled “Odds Ratio Estimates,” are obtained by exponentiating each of the parameter estimates. However, these odds ratio estimates can be misleading for continuous predictor variables or in the presence of interaction terms. For example, for continuous variables like AGE, exponentiating the estimated coefficient gives the odds ratio for a one-unit change in AGE. Also, exponentiating the estimated coefficient for CAT gives the odds ratio estimate (CAT = 1 vs. CAT = 0) for a subject whose cholesterol count is zero, which is impossible.

B. PROC GENMOD

Next, we illustrate the use of PROC GENMOD with the Evans County data. PROC GENMOD can be used to run generalized linear models (GLM) and generalized estimating equations (GEE) models, including unconditional logistic regression, which is a special case of GLM. The link function and the distribution of the outcome are specified in the model statement. LINK = LOGIT and DIST = BINOMIAL are the MODEL statement options that specify a logistic regression. Options requested in the MODEL statement are preceded by a forward slash. The code that follows runs the same model as the preceding PROC LOGISTIC:

```
PROC GENMOD DATA = REF.EVANS DESCENDING;
MODEL CHD = CAT AGE CHL ECG SMK HPT CH CC/LINK = LOGIT DIST = BINOMIAL;
ESTIMATE 'OR (CHL = 220, HPT = 1)' CAT 1 CC 220 CH 1/EXP;
ESTIMATE 'OR (CHL = 220, HPT = 0)' CAT 1 CC 220 CH 0/EXP;
CONTRAST 'LRT for interaction terms' CH 1, CC 1;
RUN;
```

The DESCENDING option in the PROC GENMOD statement instructs SAS that the outcome event of interest is CHD = 1 rather than the default, CHD = 0. An optional ESTIMATE statement can be used to obtain point estimates, confidence intervals, and a Wald test for a linear combination of parameters (e.g., $\beta_1 + 1\beta_6 + 220\beta_7$). The EXP option in the ESTIMATE statement exponentiates the requested linear combination of parameters. In this example, two odds ratios are requested using the interaction parameters:

1. $\exp(\beta_1 + 1\beta_6 + 220\beta_7)$ is the odds ratio for CAT = 1 vs. CAT = 0 for HPT = 1 and CHOL = 220
2. $\exp(\beta_1 + 0\beta_6 + 220\beta_7)$ is the odds ratio for CAT = 1 vs. CAT = 0 for HPT = 0 and CHOL = 220

The quoted text following the word ESTIMATE is a "label" that is printed in the output. The user is free to define his/her own label. The CONTRAST statement, as used in this example, requests a likelihood ratio test on the two interaction terms (CH and CC). The CONTRAST statement also requires that the user define a label. The same CONTRAST statement in PROC LOGISTIC would produce a generalized Wald test statistic, rather than a likelihood ratio test, for the two interaction terms.

The output produced from PROC GENMOD follows:

```
                    The GENMOD Procedure

                     Model Information

          Data Set                      WORK.EVANS1
          Distribution                   Binomial
          Link Function                    Logit
          Dependent Variable                chd
          Observations Used                 609

                     Response Profile

          Ordered                        Total
           Value          chd          Frequency
          1                1               71
          2                0              538

PROC GENMOD is modeling the probability that chd='1'.
```

```
                    Criteria for Assessing Goodness of Fit

     Criterion                 DF          Value           Value/DF

     Deviance                 600        347.2295           0.5787
     Scaled Deviance          600        347.2295           0.5787
     Pearson Chi-Square       600        799.0652           1.3318
     Scaled Pearson X2        600        799.0652           1.3318
     Log Likelihood                     −173.6148

Algorithm converged


                       Analysis of Parameter Estimates

                         Standard        Wald 95%          Chi-
Parameter    Estimate      Error     Confidence Limits    Square   Pr > ChiSq

Intercept     −4.0497     1.2550     −6.5095   −1.5900    10.41      0.0013
CAT          −12.6895     3.1047    −18.7746   −6.6045    16.71      <.0001
AGE            0.0350     0.0161      0.0033    0.0666     4.69      0.0303
CHL           −0.0055     0.0042     −0.0137    0.0027     1.70      0.1923
ECG            0.3671     0.3278     −0.2754    1.0096     1.25      0.2627
SMK            0.7732     0.3273      0.1318    1.4146     5.58      0.0181
HPT            1.0466     0.3316      0.3967    1.6966     9.96      0.0016
CH            −2.3318     0.7427     −3.7874   −0.8762     9.86      0.0017
CC             0.0692     0.0144      0.0410    0.0973    23.20      <.0001
Scale          1.0000     0.0000      1.0000    1.0000
```

NOTE: The scale parameter was held fixed.

```
                        Contrast Estimate Results

                       L'Beta   Standard                        Chi-    Pr >
Label                 Estimate   Error    Confidence  Limits   Square  ChiSq

Log OR (ch1 = 220,     0.1960    0.4774    −0.7397    1.1318    0.17   0.6814
  hpt = 1)
Exp(Log OR (chl=220,   1.2166    0.5808     0.4772    3.1012
  hpt=1))
Log OR (chl = 220,     2.5278    0.6286     1.2957    3.7599   16.17   <.0001
  hpt = 0)
Exp(Log OR (chl=220,  12.5262    7.8743     3.6537   42.9445
  hpt=0))

                            Contrast Results

    Contrast                      DF     Chi-Square     Pr > ChiSq     Type

    LRT for interaction terms      2        53.16          <.0001        LR
```

The table titled “Contrast Estimate Results” gives the odds ratios requested by the ESTIMATE statement. The estimated odds ratio for CAT = 1 vs. CAT = 0 for a hypertensive subject with a 220 cholesterol count is **exp**(0.1960) = 1.2166. The estimated odds ratio for CAT = 1 vs. CAT = 0 for a nonhypertensive subject with a 220 cholesterol count is **exp**(2.5278) = 12.5262. The table titled "Contrast Results" gives the chi-square test statistic (53.16) and p-value (<0.0001) for the likelihood ratio test on the two interaction terms.

C. Events/Trials Format

The Evans County dataset **evans.dat** contains individual level data. Each observation represents an individual subject. PROC LOGISTIC and PROC GENMOD also accommodate summarized binomial data in which each observation contains a count of the number of events and trials for a particular pattern of covariates. The dataset EVANS2 summarizes the 609 observations of the EVANS data into eight observations, where each observation contains a count of the number of events and trials for a particular pattern of covariates. The dataset contains five variables described below:

CASES – number of coronary heart disease cases
TOTAL – number of subjects at risk in the stratum
CAT – serum catecholamine level (1 = high, 0 = normal)
AGEGRP – dichotomized age variable (1 = age $\geq$ 55, 0 = age < 55)
ECG – electrocardiogram abnormality (1 = abnormal, 0 = normal)

The code to produce the dataset is shown next. The dataset is small enough that it can be easily entered manually.

```
DATA EVANS2;
INPUT CASES TOTAL CAT AGEGRP ECG;
CARDS;
17  274  0  0  0
15  122  0  1  0
7    59  0  0  1
5    32  0  1  1
1     8  1  0  0
9    39  1  1  0
3    17  1  0  1
14   58  1  1  1
;
```

To run a logistic regression on the summarized data EVANS2, the response is put into an *EVENTS/TRIALS* form for either PROC LOGISTIC or PROC GENMOD. The model is stated as follows:

$$\text{logit } P(\text{CHD} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{CAT} + \beta_2\text{AGEGRP} + \beta_3\text{ECG}$$

The code to run the model in PROC LOGISTIC using the dataset EVANS2 is:

```
PROC LOGISTIC DATA = EVANS2;
MODEL CASES/TOTAL = CAT AGEGRP ECG;
RUN;
```

The code to run the model in PROC GENMOD using the dataset EVANS2 is:

```
PROC GENMOD DATA = EVANS2;
MODEL CASES/TOTAL = CAT AGEGRP ECG / LINK = LOGIT DIST = BINOMIAL;
RUN;
```

The DESCENDING option is not necessary if the response is in the EVENTS/TRIALS form. The output is omitted.

The CONTRAST and ESTIMATE statements still work in PROC GENMOD when using *EVENTS/TRIALS* form. PROC LOGISTIC has a CONTRAST statement but not an ESTIMATE statement. However, linear combinations of parameter estimates can be calculated in PROC LOGISTIC using the ESTIMATE = option within the CONTRAST statement.

Suppose we wish to estimate the odds ratio comparing an individual with CAT = 1 and AGEGRP = 1 to an individual with CAT = 0 and AGEGRP = 0 controlling for ECG. The odds ratio is **exp**($\beta_1 + \beta_2$). The code to estimate this odds ratio is shown below with the CONTRAST statement:

```
PROC LOGISTIC DATA=EVANS2;
MODEL CASES/TOTAL= CAT AGEGRP ECG;
CONTRAST ‘CAT=1 AGEGRP = 1 VS CAT=0 AGEGRP=0’ CAT 1 AGEGRP 1/
ESTIMATE=EXP;
RUN;
```

The quoted text following the word CONTRAST is a “label” that is printed in the output. The user is free to define his/her own label. The ESTIMATE = EXP option estimates exp($\beta_1 + \beta_2$). If instead we used the option ESTIMATE = PARM we would estimate $\beta_1 + \beta_2$ without the exponentiation. We could also use the option ESTIMATE = BOTH to estimate the linear combination of parameters both exponentiated and not exponentiated.

A common point of confusion with the CONTRAST statement in SAS occurs with the use of commas. Consider the following code in which PROC LOGISTIC is run with two CONTRAST statements.

```
PROC LOGISTIC DATA=EVANS2;
MODEL CASES/TOTAL= CAT AGEGRP ECG;
CONTRAST ‘1 DEGREE OF FREEDOM TEST’ CAT 1 AGEGRP 1;
CONTRAST ‘2 DEGREE OF FREEDOM TEST’ CAT 1, AGEGRP 1;
RUN;
```

The first CONTRAST statement (CAT 1 AGEGRP 1) tests the hypothesis $\beta_1 + \beta_2 = 0$ while the second CONTRAST statement which contains a comma (CAT 1, AGEGRP 1) tests the hypothesis $\beta_1 = 0$ and $\beta_2 = 0$ simultaneously. These are not the same tests. If $\beta_1 + \beta_2 = 0$, it does not necessarily mean that both $\beta_1 = 0$ and $\beta_2 = 0$. The output follows:

```
                          The LOGISTIC Procedure

                  Analysis of Maximum Likelihood Estimates

                                   Standard       Wald
Parameter    DF     Estimate        Error      Chi-Square    Pr > ChiSq

Intercept     1     -2.6163        0.2123       151.8266      <.0001
CAT           1      0.6223        0.3193         3.7978      0.0513
AGEGRP        1      0.6157        0.2838         4.7050      0.0301
ECG           1      0.3620        0.2904         1.5539      0.2126

                          Contrast Test Results

Contrast                        DF      Wald Chi-Square      Pr > ChiSq
1 DEGREE OF FREEDOM TEST         1          13.2132             0.0003
2 DEGREE OF FREEDOM TEST         2          13.4142             0.0012
```

A difference between the CONTRAST statements in PROC LOGISTIC and PROC GENMOD is that with PROC LOGISTIC the default test is the WALD test while with PROC GENMOD the default test is the likelihood ratio test.

D. Using Frequency Weights

Individual level data can also be summarized using frequency counts if the variables of interest are categorical variables. The dataset EVANS3 contains the same information as EVANS2 except that each observation represents cell counts in a four-way frequency table for the variables CHD, CAT, AGEGRP, and ECG. The variable COUNT contains the frequency counts. The code that creates EVANS3 follows:

```
DATA EVANS3;
INPUT  CHD CAT  AGEGRP ECG COUNT;
CARDS;
1   0   0   0    17
0   0   0   0   257
1   0   1   0    15
0   0   1   0   107
1   0   0   1     7
0   0   0   1    52
1   0   1   1     5
0   0   1   1    27
1   1   0   0     1
0   1   0   0     7
1   1   1   0     9
0   1   1   0    30
1   1   0   1     3
0   1   0   1    14
1   1   1   1    14
0   1   1   1    44
;
```

Whereas the dataset EVANS2 contains eight data lines, the dataset EVANS3 contains sixteen data lines. The first observation of EVANS2 indicates that out of 274 subjects with CAT = 0, AGEGRP = 0, and ECG = 0, there are 17 CHD cases in the cohort. EVANS3 uses the first two observations to produce the same information. The first observation indicates that there are 17 subjects with CHD = 1, CAT = 0, AGEGRP = 0 and ECG = 0, while the second observation indicates that there are 257 subjects with CHD = 0, CAT = 0, AGEGRP = 0, and ECG = 0.

We restate the model:

$$\text{logit } P(\text{CHD} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{CAT} + \beta_2\text{AGEGRP} + \beta_3\text{ECG}$$

The code to run the model in PROC LOGISTIC using the dataset EVANS3 is:

```
PROC LOGISTIC DATA=EVANS3 DESCENDING;
MODEL CHD = CAT AGEGRP ECG;
FREQ COUNT;
RUN;
```

The FREQ statement is used to identify the variable (e.g., COUNT) in the input dataset that contains the frequency counts. The output is omitted.

The FREQ statement can also be used with PROC GENMOD. The code follows:

```
PROC GENMOD DATA=EVANS3 DESCENDING;
MODEL CHD = CAT AGEGRP ECG / LINK=LOGIT DIST=BINOMIAL;
FREQ COUNT;
RUN;
```

E. The Analyst Application

The procedures described above are run by entering the appropriate code in the Program (or Enhanced) Editor window and then submitting (i.e., running) the program. This is the approach commonly employed by SAS users. Another option for performing a logistic regression analysis in SAS is to use the Analyst Application. In this application, procedures are selected by pointing and clicking the mouse through a series of menus and dialog boxes. This is similar to the process commonly employed by SPSS users.

The Analyst Application is invoked by selecting Solutions → Analysis → Analyst from the toolbar. Once in Analyst, the permanent SAS dataset **evans.sas7bdat** can be opened into the spreadsheet. To perform a logistic regression, select Statistics → Regression → Logistic. In the dialog box, select CHD as the Dependent variable. There is an option to use a Single trial or an Events/Trials format. Next, specify which value of the outcome should be modeled using the Model Pr{} button. In this case, we wish to model the probability that CHD equals 1. Select and add the covariates to the Quantitative box. Various analysis and output options can be selected under the Model and Statistics buttons. For example, under Statistics, the covariance matrix for the parameter estimates can be requested as part of the output. Click on OK in the main dialog box to run the program. The output generated is from PROC LOGISTIC. It is omitted here as it is similar to the output previously shown. A check of the Log window in SAS shows the code that was used to run the analysis.

Conditional Logistic Regression

A conditional logistic regression is demonstrated using PROC LOGISTIC with the MI dataset. The MI dataset contains information from a study in which each of 39 cases diagnosed with myocardial infarction is matched with two controls, yielding a total of 117 subjects.

The model is stated as follows:

$$\text{logit } \text{P}(\text{CHD} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{SMK} + \beta_2\text{SBP} + \beta_3\text{ECG} + \sum_{i=1}^{38} \gamma_i V_i$$

$$V_i = \begin{cases} 1 & \text{if } i\text{th matched triplet} \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, 2, \ldots, 38$$

The model contains 42 parameters. The data contains 117 observations. The large number of parameters compared with the number of observations indicates that an unconditional logistic analysis will yield biased results.

The SAS procedure, PROC PRINT, can be used to view the MI dataset in the output window. We first run a LIBNAME statement to access the permanent SAS dataset (**mi.sas7bdat**) assuming that it is filed on the C drive:

```
LIBNAME REF ‘C:\’;

PROC PRINT DATA = REF.MI; RUN;
```

The output for the first nine observations from running the PROC PRINT follows:

```
Obs   MATCH   PERSON   MI   SMK   SBP   ECG

1       1       1      1     0    160    1
2       1       2      0     0    140    0
3       1       3      0     0    120    0
4       2       4      1     0    160    1
5       2       5      0     0    140    0
6       2       6      0     0    120    0
7       3       7      1     0    160    0
8       3       8      0     0    140    0
9       3       9      0     0    120    0
```

The matching factor is the variable MATCH which is coded 1 for a case and 0 for the two matched controls for each case.

The code to run the conditional logistic regression follows:

```
PROC LOGISTIC DATA = MI DESCENDING;
MODEL MI = SMK SBP ECG;
STRATA MATCH;
RUN;
```

The distinguishing feature in terms of SAS syntax between requesting an unconditional and conditional logistic regression is the use of the STRATA statement for the conditional logistic regression. The STRATA statement in this example declares MATCH as the stratified (or conditioned) variable.

In earlier versions of SAS, conditional logistic regression could not be run with PROC LOGISTIC. Instead, PROC PHREG was used and can still be used for conditional logistic regression. However, PROC PHREG is primarily used to run Cox proportional hazard models. Conditional logistic regression cannot be run with PROC GENMOD.

The output for the conditional logistic regression using PROC LOGISTIC follows:

```
                         The LOGISTIC Procedure

                          Conditional Analysis

                           Model Information

       Data Set                                  REF.MI
       Response Variable                           MI
       Number of Response Levels                   2
       Number of Strata                            39
       Model                                  binary logit
       Optimization Technique             Newton-Raphson ridge


                            Strata Summary

         Response          MI          Number of
          Pattern       1     0         Strata          Frequency

         1              1     2           39              117


                         Model Fit Statistics

                             Without          With
               Criterion   Covariates      Covariates

               AIC            85.692          69.491
               SC             85.692          77.777
               −2 Log L       85.692          63.491
```

```
                 Analysis of Maximum Likelihood Estimates

                                 Standard          Wald            Pr >
Parameter   DF    Estimate        Error         Chi-Square         ChiSq

SMK          1     0.7291        0.5613           1.6873          0.1940
SBP          1     0.0456        0.0152           8.9612          0.0028
ECG          1     1.5993        0.8534           3.5117          0.0609
```

```
                    Odds Ratio Estimates

                    Point            95% Wald
        Effect     Estimate      Confidence Limits

        SMK          2.073       0.690        6.228
        SBP          1.047       1.016        1.078
        ECG          4.949       0.929       26.362
```

The odds ratio estimate for SMK = 1 vs. SMK = 0 is **exp**(0.72906) = 2.073.

Obtaining ROC Curves

Next, we demonstrate how to obtain classification tables and ROC curves using PROC LOGISTIC with the knee fracture dataset. The knee fracture dataset contains information on 348 patients of which 45 actually had a knee fracture. The logistic model contains five dichotomous predictors and is stated below:

$$\text{logit } P(\text{FRACTURE} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{FLEX} + \beta_2\text{WEIGHT} + \beta_3\text{AGECAT} + \beta_4\text{HEAD} + \beta_5\text{PATELLAR}$$

The SAS code is presented below. The model statement option, PPROB =.00 TO .50 BY .05 CTABLE, requests that a classification table be added to the default output using cutpoints of predicted probabilities from 0 to 0.50 in increments of 0.05.

```
PROC LOGISTIC DATA=REF .KNEEFR DESCENDING;
MODEL FRACTURE= FLEX WEIGHT AGECAT HEAD PATELLAR / PPROB=.00 TO .50 BY
  .05 CTABLE;
RUN;
```

The output follows:

```
               Analysis of Maximum Likelihood Estimates

                            Standard      Wald
Parameter  DF  Estimate      Error     Chi-Square   Pr > ChiSq

Intercept   1   -3.4657     0.4118      70.8372       <.0001
FLEX        1    0.5277     0.3743       1.9877       0.1586
```

(continued)

```
                                  Standard       Wald
        Parameter DF  Estimate     Error     Chi-Square   Pr > ChiSq

        WEIGHT     1    1.5056   0.4093      13.5320      0.0002
        AGECAT     1    0.5560   0.3994       1.9376      0.1639
        HEAD       1    0.2183   0.3761       0.3367      0.5617
        PATELLAR   1    0.6268   0.3518       3.1746      0.0748

                Association of Predicted Probabilities and
                           Observed Responses

            Percent Concordant       71.8     Somers' D     0.489
            Percent Discordant       22.9     Gamma         0.517
            Percent Tied              5.3     Tau-a         0.111
            Pairs                   13635     c             0.745

                           Classification Table

              Correct       Incorrect                   Percentages
Prob              Non-          Non-                                              False    False
Level    Event   Event   Event  Event   Correct  Sensitivity  Specificity   POS      NEG

0.000      45        0    303      0     12.9      100.0          0.0      87.1       .
0.050      39       93    210      6     37.9       86.7         30.7      84.3      6.1
0.100      36      184    119      9     63.2       80.0         60.7      76.8      4.7
0.150      31      200    103     14     66.4       68.9         66.0      76.9      6.5
0.200      22      235     68     23     73.9       48.9         77.6      75.6      8.9
0.250      16      266     37     29     81.0       35.6         87.8      69.8      9.8
0.300       6      271     32     39     79.6       13.3         89.4      84.2     12.6
0.350       3      297      6     42     86.2        6.7         98.0      66.7     12.4
0.400       3      301      2     42     87.4        6.7         99.3      40.0     12.2
0.450       2      301      2     43     87.1        4.4         99.3      50.0     12.5
0.500       0      303      0     45     87.1        0.0        100.0        .      12.9
```

The table in the output titled "Association of Predicted Probabilities and Observed Responses" is now described. There are 45 cases and 303 noncases of knee fracture yielding 45 × 303 = 13,635 pair combinations (4th row, 1st column of output). For these pairs, 71.8% had the case with the higher predicted probability (percent concordant in the output), 22.9% had the noncase with the higher predicted probability (percent discordant in the output), and 5.3% had the same predicted probability for the case and noncase (percent tied in the output). If the percent tied is weighted as half a concordant pair then the probability of having a concordant pair rather than a discordant pair is estimated as 0.718 + 0.5(0.053) = 0.745. This is the value of the c statistic (4th row, 2nd column of output) and is the estimate of the area under the ROC plot.

The classification table uses the patients' predicted outcome probabilities obtained from the fitted logistic model to screen each patient. The probability levels (first column) are prespecified cut points requested in the model statement. For example in the third row, the cut point is 0.100. A cut point of 0.100 indicates that any patient whose predicted probability is greater than 0.100 will receive an X-ray. In other words, if a patient has a predicted probability greater than 0.100, then the patient tests positive on the screening test. Notice that if a 0.100 cut point is used (see third row), then of the 45 patients that really had a knee fracture, 36 of them are correctly classified as events and 9 are incorrectly classified as nonevents yielding a sensitivity of 0.8 or 80%.

To produce an ROC plot, first an output dataset must be created using the OUTROC= option in the MODEL statement of PROC LOGISTIC. This output dataset contains a variable representing all the predicted probabilities as well as variables representing the corresponding sensitivity and 1 − specificity. The code to create this output dataset follows:

```
PROC LOGISTIC DATA = REF .KNEEFR DESCENDING;
MODEL FRACTURE = FLEX WEIGHT AGECAT HEAD PATELLAR/OUTROC = CAT;
RUN;
```

The new dataset is called CAT (an arbitrary choice for the user). Using PROC PRINT, the first ten observations from this dataset are printed as follows:

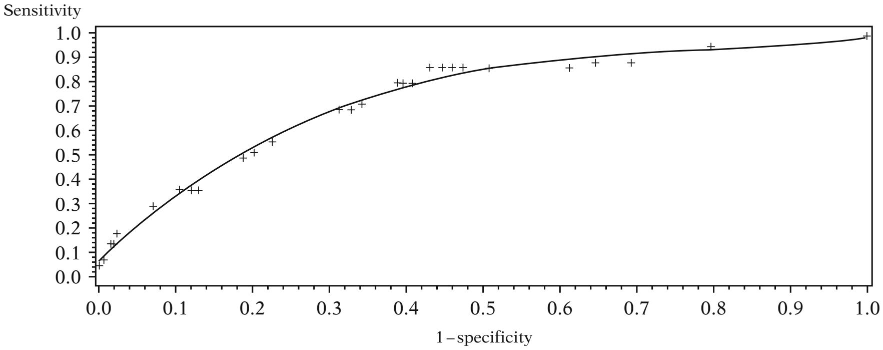
```
PROC PRINT DATA = CAT (OBS = 10); RUN;
```

```
Obs   _PROB_    _POS_   _NEG_   _FALPOS_   _FALNEG_   _SENSIT_   _1MSPEC_

1     0.49218     2      303        0          43      0.04444    0.00000
2     0.43794     3      301        2          42      0.06667    0.00660
3     0.35727     6      298        5          39      0.13333    0.01650
4     0.34116     6      297        6          39      0.13333    0.01980
5     0.31491     8      296        7          37      0.17778    0.02310
6     0.30885    13      281       22          32      0.28889    0.07261
7     0.29393    16      271       32          29      0.35556    0.10561
8     0.24694    16      266       37          29      0.35556    0.12211
9     0.23400    16      264       39          29      0.35556    0.12871
10    0.22898    22      246       57          23      0.48889    0.18812
```

The variable _PROB_ contains the predicted probabilities. The variables we wish to plot are the last two, representing the sensitivity and 1 − specificity (called _SENSIT_ and _1MSPEC_). PROC GPLOT can be used to produce a scatter plot in SAS, as shown below. The statement PLOT Y*X will plot the variable *Y* on the vertical axis and *X* on the horizontal axis. The SYMBOL statement is used before PROC GPLOT to set the plotting symbols as plus signs (VALUE=PLUS) and to plot a cubic regression to smooth the shape of the plot (INTERPOL=RC). The code and plot follow:

```
SYMBOL VALUE=PLUS INTERPOL=RC;

PROC GPLOT DATA = CAT;
PLOT_SENSIT_*_1MSPEC_;
RUN;
```

ROC Curve Using Knee Fracture Data



Polytomous Logistic Regression

A polytomous logistic regression is now demonstrated with the cancer dataset using PROC LOGISTIC. If the permanent SAS dataset **cancer.sas7bdat** is on the C drive, we can access it by running a LIBNAME statement. If the same LIBNAME statement has already been run earlier in the SAS session, it is unnecessary to rerun it.

```
LIBNAME REF ‘C:\’;
```

First a PROC PRINT is run on the cancer dataset.

```
PROC PRINT DATA = REF.CANCER; RUN;
```

The output for the first eight observations from running the proc print follows:

```
Obs    ID     GRADE   RACE   ESTROGEN   SUBTYPE   AGE   SMOKING

1     10009     1      0        0          1       0       1
2     10025     0      0        1          2       0       0
3     10038     1      0        0          1       1       0
4     10042     0      0        0          0       1       0
5     10049     0      0        1          0       0       0
6     10113     0      0        1          0       1       0
7     10131     0      0        1          2       1       0
8     10160     1      0        0          0       0       0
```

PROC LOGISTIC can be used to run a polytomous logistic regression (PROC CATMOD can also be used).

The three-category outcome variable is SUBTYPE, coded as 0 for Adenosquamous, 1 for Adenocarcinoma, and 2 for Other. The model is stated as follows:

$$\ln\left[\frac{\mathrm{P}(\mathrm{SUBTYPE}=g|\mathbf{X})}{\mathrm{P}(\mathrm{SUBTYPE}=0|\mathbf{X})}\right]=\alpha_g+\beta_{g1}\mathrm{AGE}+\beta_{g2}\mathrm{ESTROGEN}+\beta_{g3}\mathrm{SMOKING}$$

$$\text{where } g=1,2$$

By default, PROC LOGISTIC assumes the highest level of the outcome variable is the reference group. If we wish to make SUBTYPE = 0 (i.e., Adenosquamous) the reference group we can use the DESCENDING option in a similar manner as we did when we ran a standard logistic regression using PROC LOGISTIC. The code follows:

```
PROC LOGISTIC DATA = REF.CANCER DESCENDING;
MODEL SUBTYPE = AGE ESTROGEN SMOKING/LINK = GLOGIT;
RUN;
```

The key difference in the syntax for specifying a polytomous rather than a standard logistic regression using PROC LOGISTIC is the LINK = GLOGIT option in the MODEL statement. LINK = GLOGIT requests a generalized logit link function for the model. If a three (or more) level outcome is specified in the model statement without using the LINK = option, the default analysis is an ordinal logistic regression which uses a cumulative logit link function (see next section).

The output using PROC LOGISTIC for the polytomous analysis follows:

```
                        The LOGISTIC Procedure

                          Model Information

        Data Set                                REF.CANCER
        Response Variable                        SUBTYPE
        Number of Response Levels                   3
        Model                                generalized logit
        Optimization Technique                Newton-Raphson

              Number of Observations Read          288
              Number of Observations Used          286

                           Response Profile

              Ordered                          Total
               Value          SUBTYPE        Frequency

              1                  2               57
              2                  1               45
              3                  0              184

Logits modeled use SUBTYPE=0 as the reference category.


                        Model Fit Statistics

                          Intercept      Intercept and
            Criterion       Only          Covariates
            AIC           516.623          510.405
            SC            523.935          539.653
            −2 Log L      512.623          494.405
```

```
               Testing Global Null Hypothesis: BETA=0

        Test                 Chi-Square    DF     Pr > ChiSq

        Likelihood Ratio       18.2184      6       0.0057
        Score                  15.9442      6       0.0141
        Wald                   13.9422      6       0.0303


                      Type 3 Analysis of Effects

                                       Wald
        Effect          DF          Chi-Square          Pr > ChiSq

        AGE              2            5.9689              0.0506
        ESTROGEN         2            3.5145              0.1725
        SMOKING          2            6.5403              0.0380


              Analysis of Maximum Likelihood Estimates

                                        Standard       Wald
 Parameter   SUBTYPE   Estimate        Error      Chi-Square    Pr > ChiSq

 Intercept      2       -1.2032       0.3190       14.2290        0.0002
 Intercept      1       -1.8822       0.4025       21.8691        <.0001
 AGE            2        0.2823       0.3280        0.7408        0.3894
 AGE            1        0.9871       0.4118        5.7456        0.0165
 ESTROGEN       2       -0.1071       0.3067        0.1219        0.7270
 ESTROGEN       1       -0.6439       0.3436        3.5126        0.0609
 SMOKING        2       -1.7910       1.0463        2.9299        0.0870
 SMOKING        1        0.8895       0.5253        2.8666        0.0904


                         Odds Ratio Estimates

                                  Point               95% Wald
 Effect          SUBTYPE         Estimate          Confidence Limits

 AGE                2             1.326          0.697          2.522
 AGE                1             2.683          1.197          6.014
 ESTROGEN           2             0.898          0.492          1.639
 ESTROGEN           1             0.525          0.268          1.030
 SMOKING            2             0.167          0.021          1.297
 SMOKING            1             2.434          0.869          6.815
```

In the above output, there are two parameter estimates for each independent variable, as there should be for this model. Since the response variable is in descending order (see the response profile in the output), the first parameter estimate compares SUBTYPE = 2 vs. SUBTYPE = 0 and the second compares SUBTYPE = 1 vs. SUBTYPE = 0. The odds ratio for AGE = 1 vs. AGE = 0 comparing SUBTYPE = 2 vs. SUBTYPE = 0 is **exp**(0.2823) = 1.326.

PROC GENMOD does not have a generalized logit link (link = glogit), and cannot run a generalized polytomous logistic regression.

Ordinal Logistic Regression

A. PROC LOGISTIC

Ordinal logistic regression is demonstrated using the proportional odds model. Either PROC LOGISTIC or PROC GENMOD can be used to run a proportional odds model. We continue to use the cancer dataset to demonstrate this model, with the variable GRADE as the response variable. The model is stated as follows:

$$\ln\left[\frac{\text{P}(\text{GRADE} \geq g|\mathbf{X})}{\text{P}(\text{GRADE} < g|\mathbf{X})}\right] = \alpha_g + \beta_1\text{RACE} + \beta_2\text{ESTROGEN} \qquad \text{where } g = 1, 2$$

The code using PROC LOGISTIC follows:

```
PROC LOGISTIC DATA = REF.CANCER DESCENDING;
MODEL GRADE = RACE ESTROGEN;
RUN;
```

The PROC LOGISTIC output for the proportional odds model follows:

```
                         The LOGISTIC Procedure

                           Model Information

           Data Set                              REF.CANCER
           Response Variable                       grade
           Number of Response Levels                 3
           Number of Observations                   286
           Link Function                           Logit
           Optimization Technique              Fisher's scoring


                            Response Profile

               Ordered                           Total
                Value           Grade          Frequency
               1                  2                53
               2                  1               105
               3                  0               128

               Score Test for the Proportional Odds
                            Assumption

               Chi-Square          DF          Pr > ChiSq

               0.9051               2           0.6360

               Analysis of Maximum Likelihood Estimates

   Parameter  DF  Estimate  Standard Error  Chi-Square   Pr > ChiSq

  Intercept    1  -1.2744       0.2286        31.0748      <.0001
  Intercept2   1   0.5107       0.2147         5.6555      0.0174
```

```
Parameter  DF  Estimate  Standard Error  Chi-Square  Pr > ChiSq
                                                      (continued)

RACE        1    0.4270      0.2720        2.4637      0.1165
ESTROGEN    1   −0.7763      0.2493        9.6954      0.0018


                         Odds Ratio Estimates

                      Point                95% Wald
      Effect        Estimate          Confidence Limits
      RACE            1.533          0.899          2.612
      ESTROGEN        0.460          0.282          0.750
```

The Score test for the proportional odds assumption yields a chi-square value of 0.9051 and a *p*-value of 0.6360. Notice that there are two intercepts, but only one parameter estimate for each independent variable.

B. PROC GENMOD

PROC GENMOD can also be used to perform an ordinal regression; however, it does not provide a test of the proportional odds assumption. The code is as follows:

```
PROC GENMOD DATA = REF.CANCER DESCENDING;
MODEL GRADE = RACE ESTROGEN/ LINK=CUMLOGIT DIST=MULTINOMIAL;
RUN;
```

Recall that with PROC GENMOD, the link function (LINK=) and the distribution of the response variable (DIST=) must be specified. The proportional odds model uses the cumulative logit link function, while the response variable follows the multinomial distribution. The LINK = CUMLOGIT option could also be used with PROC LOGISTIC but it is unnecessary since that is the default when the response variable has three or more levels. The output is omitted.

C. Analyst Application

The Analyst Application can also be used to run an ordinal regression model. Once the **cancer.sas7bdat** dataset is opened in the spreadsheet, select Statistics → Regression → Logistic. In the dialog box, select GRADE as the Dependent variable. Next, specify which value of the outcome should be modeled using the Model Pr{} button. In this case, we wish to model the “Upper (decreasing) levels” (i.e., 2 and 1) against the lowest level. Select and add the covariates (RACE and ESTROGEN) to the Quantitative box. Various analysis and output options can be selected under the Model and Statistics buttons. For example, under Statistics, the covariance matrix for the parameter estimates can be requested as part of the output.

Click on OK in the main dialog box to run the program. The output generated is from PROC LOGISTIC and is identical to the output presented previously. A check of the Log window in SAS shows the code that was used to run the analysis.

Modeling Correlated Dichotomous Outcomes with GEE

The programming of a GEE model with the infant care dataset is demonstrated using PROC GENMOD. The model is stated as follows:

$$\text{logit P(OUTCOME} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{BIRTHWGT} + \beta_2\text{GENDER} + \beta_3\text{DIARRHEA}$$

The code and output are shown for this model assuming an AR1 correlation structure. The code for specifying other correlation structures using the REPEATED statement in PROC GENMOD is shown later in this section, although the output is omitted.

First a PROC PRINT will be run on the infant care dataset. Again, the use of the following LIBNAME statement assumes the permanent SAS dataset is stored on the C drive.

```
LIBNAME REF 'C:\';
PROC PRINT DATA = REF.INFANT; RUN;
```

The output for one infant obtained from running the PROC PRINT is presented. Each observation represents one of the nine monthly measurements.

```
IDNO    MONTH    OUTCOME    BIRTHWGT    GENDER    DIARRHEA

244       1         0         2850         2          0
244       2         1         2850         2          1
244       3         1         2850         2          0
244       4         0         2850         2          0
244       5         0         2850         2          0
244       6         0         2850         2          0
244       7         0         2850         2          0
244       8         0         2850         2          0
244       9         0         2850         2          0
```

The code for running a GEE model with an AR1 correlation structure follows:

```
PROC GENMOD DATA=REF.INFANT DESCENDING;
CLASS IDNO MONTH;
MODEL OUTCOME = BIRTHWGT GENDER DIARRHEA / DIST=BIN LINK=LOGIT;
REPEATED SUBJECT=IDNO / TYPE=AR(1) WITHIN=MONTH CORRW;
ESTIMATE 'log odds ratio (DIARRHEA 1 vs 0)' DIARRHEA 1/EXP;
CONTRAST 'Score Test BIRTHWGT and DIARRHEA' BIRTHWGT 1, DIARRHEA 1;
RUN;
```

The variable defining the cluster (infant) is IDNO. The variable defining the order of measurement within a cluster is MONTH. Both these variables must be listed in the CLASS statement. If the user wishes to have dummy variables defined from any nominal independent variables, these can also be listed in the CLASS statement.

The LINK and DIST options in the MODEL statement define the link function and the distribution of the response. Actually, for a GEE model, the distribution of the response is not specified. Rather a GEE model requires that the mean–variance relationship of the response be specified. What the DIST = BINOMIAL option does is to define the mean–variance relationship of the response to be the same as if the response followed a binomial distribution (i.e., $V(Y) = \mu\,(1 - \mu)$).

The REPEATED statement indicates that a GEE model rather than a GLM is requested. SUBJECT = IDNO in the REPEATED statement defines the cluster variable as IDNO. There are many options (following a forward slash) that can be used in the REPEATED statement. We use three of them in this example. The TYPE = AR(1) option specifies the AR1 working correlation structure, the CORRW option requests the printing of the working correlation matrix in the output window, and the WITHIN = MONTH option defines the variable (MONTH) that gives the order of measurements within a cluster. For this example, the WITHIN = MONTH option is unnecessary since the default order within a cluster is the order of observations in the data (i.e., the monthly measurements for each infant are ordered in the data from month 1 to month 9).

The ESTIMATE statement with the EXP option is used to request the odds ratio estimate for the variable DIARRHEA. The quoted text in the ESTIMATE statement is a label defined by the user for the printed output. The CONTRAST statement requests that the Score test be performed to simultaneously test the joint effects of the variable BIRTHWGT and DIARRHEA. If the REPEATED statement was omitted (i.e., defining a GLM rather than a GEE model), the same CONTRAST statement would produce a likelihood ratio test rather than a Score test. Recall the likelihood ratio test is not valid for a GEE model. A forward slash followed by the word WALD in the CONTRAST statement of PROC GENMOD requests results from a generalized Wald test rather than a Score test. The CONTRAST statement also requires a user-defined label.

The output produced by PROC GENMOD follows:

```
                             The GENMOD Procedure

                              Model Information

                   Data Set                      REF.INFANT
                   Distribution                   Binomial
                   Link Function                    Logit
                   Dependent Variable              outcome
                   Observations Used                 1203
                   Missing Values                     255


                           Class Level Information

Class   Levels                           Values

IDNO      136     00001   00002   00005   00008   00009   00010   00011   00012
                  00017   00018   00020   00022   00024   00027   00028   00030
                  00031   00032   00033   00034   00035   00038   00040   00044
                  00045   00047   00051   00053   00054   00056   00060   00061
                  00063   00067   00071   00072   00077   00078   00086   00089
                  00090   00092    ...
MONTH       9     1 2 3 4 5 6 7 8 9
```

```
                         Response Profile

                 Ordered                         Total
                  Value          Outcome       Frequency

                  1                 1               64
                  2                 0             1139

PROC GENMOD is modeling the probability that outcome=‘1’.


                 Criteria for Assessing Goodness of Fit

           Criterion                 DF         Value       Value/DF

           Deviance                 1199      490.0523       0.4087
           Scaled Deviance          1199      490.0523       0.4087
           Pearson Chi-Square       1199     1182.7485       0.9864


                 Criteria for Assessing Goodness of Fit

           Criterion               DF          Value        Value/DF

           Scaled Pearson X2      1199       1182.7485       0.9864
           Log Likelihood                    −245.0262

Algorithm converged.


                  Analysis of Initial Parameter Estimates

                               Standard        Wald 95%          Chi-      Pr >
Parameter  DF   Estimate        Error     Confidence Limits     Square     ChiSq

Intercept   1   −1.4362        0.6022    −2.6165   −0.2559     5.69      0.0171
BIRTHWGT    1   −0.0005        0.0002    −0.0008   −0.0001     7.84      0.0051
GENDER      1   −0.0453        0.2757    −0.5857    0.4950     0.03      0.8694
DIARRHEA    1    0.7764        0.4538    −0.1129    1.6658     2.93      0.0871
Scale       0    1.0000        0.0000     1.0000    1.0000
NOTE: The scale parameter was held fixed.


                          GEE Model Information

           Correlation Structure                      AR(1)
           Within-Subject Effect                 MONTH (9 levels)
           Subject Effect                       IDNO (168 levels)
           Number of Clusters                          168
           Clusters With Missing Values                 32
           Correlation Matrix Dimension                  9
           Maximum Cluster Size                          9
           Minimum Cluster Size                          0

Algorithm converged.
```

```
                             Working Correlation Matrix

       Col1    Col2    Col3    Col4    Col5    Col6    Col7    Col8    Col9

Row1 1.0000  0.5254  0.2760  0.1450  0.0762  0.0400  0.0210  0.0110  0.0058
Row2 0.5254  1.0000  0.5254  0.2760  0.1450  0.0762  0.0400  0.0210  0.0110
Row3 0.2760  0.5254  1.0000  0.5254  0.2760  0.1450  0.0762  0.0400  0.0210
Row4 0.1450  0.2760  0.5254  1.0000  0.5254  0.2760  0.1450  0.0762  0.0400
Row5 0.0762  0.1450  0.2760  0.5254  1.0000  0.5254  0.2760  0.1450  0.0762
Row6 0.0400  0.0762  0.1450  0.2760  0.5254  1.0000  0.5254  0.2760  0.1450
Row7 0.0210  0.0400  0.0762  0.1450  0.2760  0.5254  1.0000  0.5254  0.2760
Row8 0.0110  0.0210  0.0400  0.0762  0.1450  0.2760  0.5254  1.0000  0.5254
Row9 0.0058  0.0110  0.0210  0.0400  0.0762  0.1450  0.2760  0.5254  1.0000


                      Analysis of GEE Parameter Estimates
                       Empirical Standard Error Estimates

Parameter  Estimate  Standard Error  95% Confidence Limits      Z    Pr > |Z|

Intercept  −1.3978       1.1960        −3.7418     0.9463   −1.17  0.2425
BIRTHWGT   −0.0005       0.0003        −0.0011     0.0001   −1.61  0.1080
GENDER      0.0024       0.5546        −1.0846     1.0894    0.00  0.9965
DIARRHEA    0.2214       0.8558        −1.4559     1.8988    0.26  0.7958


                            Contrast Estimate Results

                               Standard  95% Confidence    Chi-
Label                  Estimate   Error        Limits       Square  Pr>ChiSq

log odds ratio           0.2214   0.8558  −1.4559 1.8988     0.07    0.7958
  (DIARRHEA 1 vs 0)
Exp(log odds ratio       1.2479   1.0679   0.2332 6.6779
  (DIARRHEA 1 vs 0))


                        Contrast Results for GEE Analysis

Contrast                           DF    Chi-Square    Pr > ChiSq    Type
Score Test BIRTHWGT and DIARRHEA    2       1.93         0.3819      Score
```

The output includes a table containing “Analysis of Initial Parameter Estimates.” The initial parameter estimates are the estimates obtained from running a standard logistic regression assuming an independent correlation structure. The parameter estimation for the standard logistic regression is used as a numerical starting point for obtaining GEE parameter estimates.

Tables for GEE model information, the working correlation matrix, and GEE parameter estimates follow the initial parameter estimates in the output. Here, the working correlation matrix is a 9 × 9 matrix with an AR1 correlation structure. The table containing the GEE parameter estimates includes the empirical standard errors. Model-based standard errors could also have been requested using the MODELSE option in the REPEATED statement. The table titled "Contrast Estimate Results" contains the output requested by the ESTIMATE statement. The odds ratio estimate for DIARRHEA = 1 vs. DIARRHEA = 0 is given as 1.2479. The table titled "Contrast Results for GEE Analysis" contains the output requested by the CONTRAST statement. The *p*-value for the requested Score test is 0.3819.

Other correlation structures could be requested using the TYPE = option in the REPEATED statement. Examples of code requesting an independent, an exchangeable, a stationary 4-dependent, and an unstructured correlation structure using the variable IDNO as the cluster variable are given below.

```
REPEATED SUBJECT=IDNO / TYPE=IND;
REPEATED SUBJECT=IDNO / TYPE=EXCH;
REPEATED SUBJECT=IDNO / TYPE=MDEP(4);
REPEATED SUBJECT=IDNO / TYPE=UNSTR MAXITER=1000;
```

The ALR approach, which was described in Chap. 16, is an alternative to the GEE approach with dichotomous outcomes. It is requested by using the LOGOR = option rather than the TYPE = option in the REPEATED statement. The code requesting the alternating logistic regression (ALR) algorithm with an exchangeable odds ratio structure is:

```
REPEATED SUBJECT = IDNO / LOGOR = EXCH;
```

The MAXITER = option in the REPEATED statement can be used when the default number of 50 iterations is not sufficient to achieve numerical convergence of the parameter estimates. It is important that you make sure the numerical algorithm converged correctly to preclude reporting spurious results. In fact, the ALR model in this example, requested by the LOGOR = EXCH option, does not converge for the infant care dataset no matter how many iterations are allowed for convergence. The GEE model, using the unstructured correlation structure, also did not converge, even with MAXITER set to 1,000 iterations.

Generalized Linear Mixed Models with Dichotomous Outcomes

Generalized linear mixed models (GLMMs) can be run in SAS using PROC GLIMMIX or PROC NLMIXED. Our focus here is to illustrate PROC GLIMMIX. GLMMs are a generalization of linear mixed models in that they allow for the inclusion of fixed and random effects with nonnormally distributed outcome data. PROC GLIMMIX is a flexible procedure that can run relatively simple or quite complex models. We begin our illustration of PROC GLIMMIX by demonstrating how a standard logistic regression is run using the Evans County data. Typically, we would not use PROC GLIMMIX to run a standard logistic regression, but for illustration we present it here as a starting point for building more complicated models. The model is the same as used earlier in this appendix and is repeated below:

$$\text{logit P(CHD} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{CAT} + \beta_2\text{AGE} + \beta_3\text{CHL} + \beta_4\text{ECG} + \beta_5\text{SMK} + \beta_6\text{HPT} + \beta_7\text{CH} + \beta_8\text{CC}$$

For comparison, we show how to run this model using PROC GENMOD and then using PROC GLIMMIX. Two ESTIMATE statements are used to request odds ratio estimates derived from a linear combination of parameters and a CONTRAST statement is used to request a chunk test for the two interaction terms using the generalized Wald test. The default test with the CONTRAST statement in PROC GENMOD is the likelihood ratio test (shown earlier), but because the CONTRAST statement in PROC GLIMMIX does not produce a likelihood ratio test statistic, the WALD option was used in PROC GENMOD for comparability. The CHISQ option in the CONTRAST statement of PROC GLIMMIX requests a Wald chi-square test statistic to be given in addition to the default F test. The code is shown below:

```
PROC GENMOD DATA=REF.EVANS DESCENDING;
MODEL CHD = CAT AGE CHL ECG SMK HPT CH CC/LINK=LOGIT DIST=BINOMIAL;
ESTIMATE 'OR (CHL=220, HPT=1)' CAT 1 CC 220 CH 1/EXP;
ESTIMATE 'OR (CHL=220, HPT=0)' CAT 1 CC 220 CH 0/EXP;
CONTRAST 'WALD test for interaction terms' CH 1, CC 1/WALD;
RUN;

PROC GLIMMIX DATA = REF.EVANS;
MODEL CHD = CAT AGE CHL ECG SMK HPT CH CC/DIST=BIN LINK=LOGIT SOLUTION
NOSCALE DDFM=NONE;
ESTIMATE 'OR (CHL=220, HPT=1)' CAT 1 CC 220 CH 1/EXP;
ESTIMATE 'OR (CHL=220, HPT=0)' CAT 1 CC 220 CH 0/EXP;
CONTRAST 'Wald test for interaction terms' CH 1, CC 1/CHISQ;
RUN;
```

Notice that PROC GLIMMIX does not use the DESCENDING option (as does PROC GENMOD) to indicate that CHD = 1 is the value for an event rather than CHD = 0. Both procedures use the DIST = BIN and LINK = LOGIT in the MODEL statement to indicate that the outcome follows a binomial distribution with a logit link function. The SOLUTION option in PROC GLIMMIX requests that the parameter estimates for the fixed effects appear in the output. Parameter estimates are given with PROC GENMOD by default. The NOSCALE and DDFM=NONE options in PROC GLIMMIX allow the test of significance (using a T test in PROC GLIMMIX) to be equivalent to that given by PROC GENMOD (using a chi-square test). The ESTIMATE statements do not differ in the two procedures.

The output is essentially the same as that given earlier in this appendix when PROC GENMOD was first described and is omitted here.

PROC GLIMMIX can be used to run GEE models that allow random effects as well as fixed effects. In the previous section, a GEE model was run using PROC GENMOD on the infant dataset. We now consider the same model and demonstrate how PROC

GLIMMIX can be used for GEE models. The model is restated below:

$$\text{logit P(OUTCOME} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{BIRTHWGT} + \beta_2\text{GENDER} + \beta_3\text{DIARRHEA}$$

To illustrate how a GEE model is run using PROC GLIMMIX, we compare the code used to run PROC GENMOD for the same model described in the previous section on GEE. This model only contains fixed effects. The code is shown first for PROC GENMOD and then for PROC GLIMMIX. An AR1 structure is chosen as the working correlation matrix in PROC GENMOD and as the working covariance matrix in PROC GLIMMIX. The code follows:

```
PROC GENMOD DATA=REF.INFANT DESCENDING;
CLASS IDNO MONTH;
MODEL OUTCOME = BIRTHWGT GENDER DIARRHEA /DIST=BIN LINK=LOGIT;
REPEATED SUBJECT=IDNO / TYPE=AR(1) WITHIN=MONTH CORRW;
ESTIMATE ‘log odds ratio (DIARRHEA 1 vs 0)’ DIARRHEA 1/EXP;
RUN;

PROC GLIMMIX DATA=REF.INFANT EMPIRICAL;
CLASS IDNO MONTH;
MODEL OUTCOME = BIRTHWGT GENDER DIARRHEA /DIST=BIN LINK=LOGIT SOLUTION
CHISQ;
RANDOM _RESIDUAL_ / SUBJECT=IDNO TYPE=AR(1) VCORR;
ESTIMATE ‘log odds ratio (DIARRHEA 1 vs 0)’ DIARRHEA 1/EXP;
RUN;
```

The DESCENDING option is not used in PROC GLIMMIX as it recognizes the value OUTCOME = 1 as an event and OUTCOME = 0 as a nonevent. The EMPIRICAL option in the PROC GLIMMIX statement requests empirical standard errors for the parameter estimates. The options shown in the MODEL statement of PROC GLIMMIX were described in the previous example. The RANDOM statement with the key word _RESIDUAL_ following it plays the same role in PROC GLIMMIX as the REPEATED statement does in PROC GENMOD. The cluster variable in the infant dataset, IDNO, is defined with the SUBJECT = option. The TYPE = option defines the correlation structure for the residuals (the R matrix). The VCORR option requests that the correlation matrix for the random error be printed in the output. The output from PROC GLIMMIX follows:

```
                        The GLIMMIX Procedure

                          Model Information

          Data Set                          REF.INFANT
          Response Variable                 OUTCOME
          Response Distribution             Binomial
          Link Function                     Logit
          Variance Function                 Default
          Variance Matrix Blocked By        IDNO
          Estimation Technique              Residual PL
          Degrees of Freedom Method         Between-Within
          Fixed Effects SE Adjustment       Sandwich – Classical
```

```
                  Class Level Information

Class      Levels                  Values

IDNO        136       1 2 5 8 9 10 11 12 17 18 20 22 24 27 28 30 31
                      32 33 34 35 38 40 44 45 47 51 53 54 56 60 61
                      63 67 71 72 77 78 86 89 90 92 94 102 111 112
                      166 167 174 175 176 178 181 183 192 193 194
                      195 196 197 199 202 204 205 207 208 216 218
                      219 221 222 223 227 230 232 237 241 242 244
                      245 248 249 250 252 253 254 255 262 263 268
                      269 276 277 278 279 281 282 283 284 287 288
                      289 290 291 293 295 298 299 300 301 306 309
                      310 315 318 319 321 324 326 330 331 332 334
                      335 337 338 339 340 341 344 346 347 349 351
                      354 355
MONTH         9       1 2 3 4 5 6 7 8 9


           Number of Observations Read         1458
           Number of Observations Used         1203

                        Dimensions

           R-side Cov. Parameters                 2
           Columns in X                           4
           Columns in Z per Subject               0
           Subjects (Blocks in V)               136
           Max Obs per Subject                    9

                  Optimization Information

                                     Dual Quasi-
           Optimization Technique       Newton

         Parameters in                   1
           Optimization
         Lower Boundaries                1
         Upper Boundaries                1

                  Optimization Information

           Fixed Effects                   Profiled
           Residual Variance               Profiled
           Starting From                     Data

                      Fit Statistics

       −2 Res Log Pseudo-Likelihood             6668.60
       Generalized Chi-Square                   1164.06
       Gener. Chi-Square / DF                      0.97
```

```
                         Estimated V Correlation Matrix for IDNO 1

Row   Col1      Col2     Col3     Col4     Col5     Col6     Col7     Col8     Col9

1    1.0000    0.5370   0.2884   0.1549   0.08317  0.04467  0.02399  0.01288  0.006918
2    0.5370    1.0000   0.5370   0.2884   0.1549   0.08317  0.04467  0.02399  0.01288
3    0.2884    0.5370   1.0000   0.5370   0.2884   0.1549   0.08317  0.04467  0.02399
4    0.1549    0.2884   0.5370   1.0000   0.5370   0.2884   0.1549   0.08317  0.04467
5    0.08317   0.1549   0.2884   0.5370   1.0000   0.5370   0.2884   0.1549   0.08317
6    0.04467   0.08317  0.1549   0.2884   0.5370   1.0000   0.5370   0.2884   0.1549
7    0.02399   0.04467  0.08317  0.1549   0.2884   0.5370   1.0000   0.5370   0.2884
8    0.01288   0.02399  0.04467  0.08317  0.1549   0.2884   0.5370   1.0000   0.5370
9    0.006918  0.01288  0.02399  0.04467  0.08317  0.1549   0.2884   0.5370   1.0000

                                  The GLIMMIX Procedure

                             Covariance Parameter Estimates

                   Cov Parm     Subject     Estimate     Standard Error

                   AR(1)         IDNO        0.5370         0.02514
                   Residual                  0.9709         0.05150

                               Solutions for Fixed Effects

                 Effect      Estimate   Standard Error   DF    t Value  Pr > |t|

                 Intercept  −1.3969       1.1949         133   −1.17    0.2445
                 BIRTHWGT   −0.00049      0.000307       133   −1.61    0.1095
                 GENDER      0.004201     0.5549         133    0.01    0.9940
                 DIARRHEA    0.2112       0.8648        1066    0.24    0.8071

                             Type III Tests of Fixed Effects

                 Effect         Num DF       Den DF       F Value       Pr > F

                 BIRTHWGT         1           133          2.60         0.1095
                 GENDER           1           133          0.00         0.9940
                 DIARRHEA         1          1066          0.06         0.8071

                                       Estimates

                               Standard                                 Exponentiated
   Label           Estimate      Error      DF     t Value   Pr > |t|      Estimate

   log odds ratio   0.2112      0.8648     1066     0.24     0.8071        1.2352
     (DIARRHEA 1
     vs 0)
```

The model results from PROC GLIMMIX are close but not exactly the same as was obtained from PROC GENMOD in the previous section. The odds ratio estimate for DIARRHEA = 1 vs. DIARRHEA = 0 derived from the ESTIMATE statement is given as 1.2352 (last row at the end of the output). The odds ratio was estimated at 1.2479 using PROC GENMOD.

The default optimization method for parameter estimation in PROC GLIMMIX is a residual pseudo-likelihood technique that utilizes a working *covariance* structure provided by the user. The AR(1) covariance structure contains two parameters: a correlation parameter (estimated at 0.5370 in the output) and a variance parameter (estimated at 0.9709 in the output). The AR1 correlation parameter using PROC GENMOD was estimated at 0.5254 (slightly different form 0.5370 with PROC GLIMMIX). PROC GLIMMIX provides F test statistics (or equivalent T statistics) rather than chi-square Wald chi-square statistics for the parameter estimates in the default output. The CHISQ option in the MODEL statement will additionally add chi-square test statistics in the output.

The output from PROC GLIMMIX uses the terminology R-side parameters for covariance parameters of the residual matrix (R matrix) and G-side parameters for the covariance parameters of the random effects (G matrix).

The next example demonstrates how to run a model containing a random intercept for each subject. The model, shown below, assumes an R matrix with independent correlation structure and a scalar (1 × 1) G matrix:

$$\begin{aligned}\text{logit P}(\text{OUTCOME} = 1|\mathbf{X}) = {} & (\beta_0 + b_{0i}) + \beta_1\text{BIRTHWGT} + \beta_2\text{GENDER} \\ & + \beta_3\text{DIARRHEA} \\ & \text{where } b_{0i} \text{ represents the random effect for subject} \\ & i \text{ and is normally distributed with mean} = 0 \\ & \text{and variance} = \sigma_s{}^2, \text{ i.e.,} b_{0i} \sim \text{N}(0, \sigma_s{}^2)\end{aligned}$$

The code to run this model in PROC GLIMMIX follows:

```
PROC GLIMMIX DATA=REF.INFANT;
CLASS IDNO;
MODEL OUTCOME = BIRTHWGT GENDER DIARRHEA / DIST = BIN LINK = LOGIT SOLUTION;
RANDOM INTERCEPT / SUBJECT = IDNO;
RANDOM _RESIDUAL_;
RUN;
```

The first RANDOM statement is followed by the key word INTERCEPT. The SUBJECT = option specifies the variable IDNO as the cluster variable. The second RANDOM statement (RANDOM _RESIDUAL_) is optional and requests variance estimates for the residual in the output but also provides parameter estimates identical to those provided by the SAS macro called GLIMMIX (a precursor of the GLIMMIX procedure). The output is omitted.

The next model includes a random slope for the variable DIARRHEA in addition to the random intercept. The model is stated as follows:

$$\begin{aligned}\text{logit P}(\text{OUTCOME} = 1|\mathbf{X}) = {} & (\beta_0 + b_{0i}) + \beta_1\text{BIRTHWGT} + \beta_2\text{GENDER} \\ & + (\beta_3 + b_{3i})\text{DIARRHEA} \\ & \text{where } b_{0i} \text{ represents the random intercept for} \\ & \text{subject } i\text{, and where } b_{3i} \text{ represents a random} \\ & \text{slope with the variable DIARRHEA} \\ & \text{for subject } i\text{, } b_{0i} \sim \text{N}(0, \sigma_s{}^2) \text{ and } b_{3i} \sim \text{N}(0, \sigma_D{}^2)\end{aligned}$$

The code to run this model with two random effects follows:

```
PROC GLIMMIX DATA = REF.INFANT;
CLASS IDNO;
MODEL OUTCOME = BIRTHWGT GENDER DIARRHEA / DIST=BIN LINK=LOGIT
  SOLUTION;
RANDOM INTERCEPT DIARRHEA / SUBJECT=IDNO TYPE=UN GCORR;
RANDOM _RESIDUAL_;
ESTIMATE ‘log odds ratio (DIARRHEA 1 vs 0)’ DIARRHEA 1/EXP;
RUN;
```

The random effects (INTERCEPT and DIARRHEA) are listed after the key word RANDOM. Since there is more than one random effect, we need to consider the covariation between the random effects. The TYPE=UN option requests an unstructured covariance structure for the working covariance structure for the random effects (here a 2×2 G matrix). The GCORR option in the first RANDOM statement requests that the correlation matrix for the G matrix be printed in the output. The RANDOM _RESIDUAL_ statements request variance estimates for the residual (the R matrix) in the output with its standard error. The SUBJECT=IDNO identifies IDNO as the cluster variable. The output follows:

```
                        The GLIMMIX Procedure

                          Model Information

         Data Set                                  REF.INFANT
         Response Variable                          OUTCOME
         Response Distribution                      Binomial
         Link Function                               Logit
         Variance Function                          Default
         Variance Matrix Blocked By                   IDNO
         Estimation Technique                      Residual PL
         Degrees of Freedom Method                 Containment

                             Dimensions

           G-side Cov. Parameters                           3
           R-side Cov. Parameters                           1
           Columns in X                                     4
           Columns in Z per Subject                         2
           Subjects (Blocks in V)                         136
           Max Obs per Subject                              9

                   Estimated G Correlation Matrix

           Effect         Row       Col1          Col2

           Intercept       1       1.0000       -0.2716
           DIARRHEA        2      -0.2716        1.0000
```

```
                  Covariance Parameter Estimates

                                                  Standard
            Cov Parm         Subject    Estimate    Error

            UN(1, 1)          IDNO        8.4913    1.3877
            UN(2, 1)          IDNO       -2.5732    2.7473
            UN(2, 2)          IDNO       10.5707    4.6568
            Residual (VC)                 0.1578    0.006670

                   Solutions for Fixed Effects

                                Standard
    Effect       Estimate        Error        DF      t Value     Pr > |t|

    Intercept    -4.4899       1.7238        133      -2.60       0.0102
    BIRTHWGT     -0.00046      0.000487     1038      -0.94       0.3453
    GENDER        0.3695       0.6908       1038       0.53       0.5928
    DIARRHEA      0.6999       0.9438         28       0.74       0.4645

                                Estimates

                            Standard                             Exponentiated
Label               Estimate   Error    DF  t Value   Pr > |t|      Estimate

log odds ratio       0.6999    0.9438   28   0.74     0.4645         2.0135
  (DIARRHEA 1 vs 0)
```

There are three G-side Covariance parameters estimated: the variance of the random intercept, the variance of the random slope for DIARRHEA, and the covariance of the random intercept and random slope. These estimates are in the table under the heading “Covariance Parameter Estimates” and labeled by UN(1, 1), UN(2, 1), and UN(2, 2) (the row and column of the unstructured G matrix). The estimated correlation between the random intercept and slope is given at −0.4256 (under the heading called “Estimated G Correlation Matrix”). There is also one R-side variance parameter estimated (obtained when using the RANDOM_RESIDUAL_ statement).

The odds ratio estimate for DIARRHEA = 1 vs. DIARRHEA = 0, requested with the ESTIMATE statement, is given as 2.0135 (last row at the end of the output). The interpretation of this odds ratio is tricky because there is a random slope component. The odds ratio for DIARRHEA for the ith subject is **exp(**$(\beta_3 + b_{3i})$, where b_{3i} follows a normal distribution of mean 0 and variance σ_0^2. The interpretation of β_3 is the average log odds ratio among all subjects.

Finally, we examine some complicated variations of the previous model. We show the code but omit the output as with this data there were issues of model stability and numerical convergence.

The previous model contained a random intercept and random slope as random effects. The following code runs the same model but adds an autoregressive AR(1) covariance structure for the residuals grouped by gender:

```
PROC GLIMMIX DATA=REF.INFANT;
CLASS IDNO;
MODEL OUTCOME = BIRTHWGT GENDER DIARRHEA / DIST=BIN LINK=LOGIT
  SOLUTION;
RANDOM INTERCEPT DIARRHEA / SUBJECT=IDNO TYPE=UN GCORR;
RANDOM _RESIDUAL_ / SUBJECT=IDNO TYPE=AR(1) GROUP=GENDER VCORR;
RUN;
```

Here, there are two RANDOM statements, one for specifying the G matrix and the other for the residuals (the R matrix). The GROUP = GENDER option in the second RANDOM statement requests a different set of AR(1) parameters to be estimated for boy and girl infants. Typically, when a user specifies a covariance or correlation structure, the values of the covariance parameters are assumed to be the same for each cluster (subject) in the dataset. The GROUP = option allows a different set of covariance parameters to be estimated for specified subgroups.

PROC GLIMMIX also accommodates models with nested effects. As a hypothetical example, suppose 30 daycare centers were randomly sampled and within each daycare center 10 infants were sampled yielding 300 infants in all (30 × 10). Also, each infant has monthly measurements over a 9-month period. In this setting, we can consider three types of independent variables: (1) a variable like DIARRHEA whose status may vary within an infant from month-to-month, (2) a variable like GENDER which is fixed at the infant level (does not vary month-to-month), and (3) a variable that is fixed at the daycare level such as the size of the daycare center. Here we have a cluster of daycare centers and nested within each daycare center is a cluster of infants. In the infant dataset, the variable identifying each infant is called IDNO. Suppose the variable identifying the daycare center was called DAYCARE (this variable does not actually exist in the infant dataset). Consider a model with a random intercept for each infant as well as a random intercept for each daycare center. We continue to use BIRTHWEIGHT, GENDER, and DIARRHEA as fixed effects. The code to run such a model in PROC GLIMIX is:

```
PROC GLIMMIX DATA=REF.INFANT;
CLASS IDNO DAYCARE;
MODEL OUTCOME = BIRTHWGT GENDER DIARRHEA / DIST=BIN LINK=LOGIT
  SOLUTION;
RANDOM INTERCEPT/ SUBJECT=IDNO;
RANDOM INTERCEPT/ SUBJECT=DAYCARE(IDNO);
RUN;
```

The second RANDOM statement contains the option SUBJECT = DAYCARE(IDNO) which indicates that infants (IDNO) are nested within daycare centers. A random slope parameter could be added to either RANDOM statement depending on whether the slope is modeled to randomly vary by infant or by daycare center.

The SAS section of this appendix is completed. Next, modeling with SPSS software is illustrated.

SPSS

Analyses are carried out in SPSS by using the appropriate SPSS procedure on an SPSS dataset. Most users will select procedures by pointing and clicking the mouse through a series of menus and dialog boxes. The code, or command syntax, generated by these steps can be viewed (and edited by more experienced SPSS users) and is presented here for comparison to the corresponding SAS code.

The following five SPSS procedures are demonstrated:

LOGISTIC REGRESSION – This procedure is used to run a standard logistic regression.
NOMREG – This procedure is used to run a standard (binary) or polytomous logistic regression.
PLUM – This procedure is used to run an ordinal regression.
COXREG – This procedure may be used to run a conditional logistic regression for the special case in which there is only *one case per stratum*, with one (or more) controls.
GENLIN – This procedure is used to run GLM or GEE models.

SPSS does not perform generalized linear mixed models for correlated data in version 16.0.

Unconditional Logistic Regression

The first illustration presented is an unconditional logistic regression using the Evans County dataset. As discussed in the previous section, the dichotomous outcome variable is CHD and the covariates are: CAT, AGE, CHL, ECG, SMK, and HPT. Two interaction terms, CH and CC are also included. CH is the product: CAT × HPT, while CC is the product: CAT × CHL. The variables representing the interaction terms have already been included in the SPSS dataset **evans.sav**.

The model is restated as follows:

$$\begin{aligned}\text{logit P}(\text{CHD} = 1|\mathbf{X}) = &\beta_0 + \beta_1\text{CAT} + \beta_2\text{AGE} + \beta_3\text{CHL} + \beta_4\text{ECG} + \beta_5\text{SMK} \\ &+ \beta_6\text{HPT} + \beta_7\text{CH} + \beta_8\text{CC}\end{aligned}$$

The first step is to open the SPSS dataset, **evans.sav**, into the Data Editor window. The corresponding command syntax to open the file from the C drive is:

```
GET
FILE= ‘C:\evans.sav’.
```

There are three procedures that can be used to fit a standard (binary) logistic regression model: LOGISTIC REGRESSION, NOMREG, or GENLIN. The LOGISTIC REGRESSION procedure performs a standard logistic regression for a dichotomous outcome, while the NOMREG procedure can be used for dichotomous or polytomous outcomes. The GENLIN procedure can be used to run generalized linear models, including a standard logistic regression model.

To run the LOGISTIC REGRESSION procedure, select Analyze → Regression → Binary Logistic from the drop-down menus to reach the dialog box to specify the logistic model. Select CHD from the variable list and enter it into the Dependent Variable box, then select and enter the covariates into the Covariate(s) box. The default method is Enter, which runs the model with the covariates the user entered into the Covariate(s) box. Click on OK to run the model. The output generated will appear in the SPSS Viewer window.

The corresponding syntax, with the default specifications regarding the modeling process, is:

```
LOGISTIC REGRESSION VAR=chd
  /METHOD=ENTER cat age ch1 ecg smk hpt ch cc
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

To obtain 95% confidence intervals for the odds ratios, before clicking on OK to run the model, select the PASTE button in the dialog box. A new box appears which contains the syntax shown above. Insert /PRINT=CI(95) before the /CRITERIA line as shown below:

```
LOGISTIC REGRESSION VAR=chd
  /METHOD=ENTER cat age ch1 ecg smk hpt ch cc
  /PRINT=CI(95)
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

Then click on OK to run the model.

The LOGISTIC REGRESSION procedure models the P(CHD = 1) rather than P(CHD = 0) by default. The internal coding can be checked by examining the table "Dependent Variable Encoding."

The output produced by LOGISTIC REGRESSION follows:

Logistic Regression

Case Processing Summary

Unweighted cases[a]		*N*	Percent
Selected cases	Included in analysis	609	100.0
	Missing cases	0	.0
	Total	609	100.0
Unselected cases		0	.0
Total		609	100.0

[a] If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original value	Internal value
.00	0
1.00	1

Model Summary

Step	−2 Log likelihood	Cox & Snell R square	Nagelkerke R square
1	347.230	.139	.271

Variables in the Equation

								95.0% C.I. for EXP(*B*)	
		B	S.E.	Wald	df	Sig.	Exp (*B*)	Lower	Upper
Step 1[a]	CAT	−12.688	3.104	16.705	1	.000	.000	.000	.001
	AGE	.035	.016	4.694	1	.030	1.036	1.003	1.069
	CHL	−.005	.004	1.700	1	.192	.995	.986	1.003
	ECG	.367	.328	1.254	1	.263	1.444	.759	2.745
	SMK	.773	.327	5.582	1	.018	2.167	1.141	4.115
	HPT	1.047	.332	9.960	1	.002	2.848	1.487	5.456
	CH	−2.332	.743	9.858	1	.002	.097	.023	.416
	CC	.069	.014	23.202	1	.000	1.072	1.042	1.102
	Constant	−4.050	1.255	10.413	1	.001	.017		

[a] Variable(s) entered on step 1: CAT, AGE, CHL, ECG, SMK, HPT, CH, CC.

The estimated coefficients for each variable (labeled B) and their standard errors, along with the Wald chi-square test statistics and corresponding *p*-values, are given in the table titled "Variables in the Equation." The intercept is labeled "Constant" and is given in the last row of the table. The odds ratio estimates are labeled EXP(B) in the table, and are obtained by exponentiating the corresponding coefficients. As noted previously in the SAS section, these odds ratio estimates can be misleading for continuous variables or in the presence of interaction terms.

The negative 2 log likelihood statistic for the model, 347.23, is presented in the table titled "Model Summary." A likelihood ratio test statistic to asses the significance of the two interaction terms can be performed by running a no-interaction model and subtracting the negative 2 log likelihood statistic for the current model from that of the no-interaction model.

Suppose we wish to estimate the odds ratio for CAT $= 1$ vs. CAT $= 0$ among those with HPT $= 0$ and CHOL $= 220$. This odds ratio is **exp**$(\beta_1 + 220\beta_8)$. From the output, this is estimated at **exp(**$-12.688 + 220 \times .069$). This is an example of an odds ratio ascertained as a linear combination of parameters. Obtaining a linear combination of parameter estimates along with the corresponding standard error and 95% confidence interval is not straightforward in SPSS as it is in SAS (with an ESTIMATE statement) or in Stata (with the LINCOM command). However, there is a way to "trick" SPSS into doing this. Since, in this example, we are interested in estimating the odds ratio for CAT among those who have a cholesterol level of 220 (CHL $=$ 220), the trick is to create a new variable for cholesterol such that when the cholesterol level is 220, the new variable takes the value zero. For that situation the parameter for the product term will drop out in the calculation of the odds ratio. The new variable we shall create will be called CHL220 and be equal to CHL minus 220. We shall also create a product term CAT × CHL220. This can be accomplished using the dialog box: Transform → Compute Variable and then defining the new variable or by using the following syntax:

```
COMPUTE chl220=chl-220.
EXECUTE.

COMPUTE cc220= cat * chl220.
EXECUTE.
```

Now run the same model as before, except replace CHL220 for CHL and CC220 for the product term CC. The desired odds ratio will be just **exp(**β_1). The syntax is as follows:

```
LOGISTIC REGRESSION VAR=chd
  /METHOD=ENTER cat age chl220 ecg smk hpt ch cc220
  /PRINT=CI(95)
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

The output containing the parameter estimate follows:

Variables in the Equation

							95.0% C.I. for EXP(*B*)	
	B	S.E.	Wald	df	Sig.	Exp(*B*)	Lower	Upper
Step 1 cat	2.528	.629	16.170	1	.000	12.526	3.654	42.944
age	.035	.016	4.694	1	.030	1.036	1.003	1.069
chl220	−.005	.004	1.700	1	.192	.995	.986	1.003
ecg	.367	.328	1.254	1	.263	1.444	.759	2.745
smk	.773	.327	5.582	1	.018	2.167	1.141	4.115
hpt	1.047	.332	9.961	1	.002	2.848	1.487	5.456
ch	−2.332	.743	9.858	1	.002	.097	.023	.416
cc220	.069	.014	23.202	1	.000	1.072	1.042	1.102
Constant	−5.250	.960	29.906	1	.000	.005		

The first row of the output shows that the estimated odds ratio for CAT = 1 vs. CAT = 0 among those with HPT = 0 and CHOL = 220 using this new coding is **exp(**2.528) = 12.526 with corresponding 95% confidence interval (3.654, 42.944).

With the NOMREG procedure, the values of the outcome are sorted in ascending order with the last (or highest) level of the outcome variable as the reference group. If we wish to model P(CHD = 1), as was done in the previous analysis with the LOGISTIC REGRESSION procedure, the variable CHD must first be recoded so that CHD = 0 is the reference group. This process can be accomplished using the dialog boxes. The command syntax to recode CHD into a new variable called NEWCHD is:

```
RECODE
  chd
  (1=0) (0=1) INTO newchd.
EXECUTE.
```

To run the NOMREG procedure, select Analyze → Regression → Multinomial Logistic from the drop-down menus to reach the dialog box to specify the logistic model. Select NEWCHD from the variable list and enter it into the Dependent Variable box, then select and enter the covariates into the Covariate(s) box. The default settings in the Model dialog box are "Main Effects" and "Include intercept in model." With the NOMREG procedure, the covariance matrix can be requested as part of the model statistics. Click on the Statistics button and check "Asymptotic covariances of parameter estimates" to include a covariance matrix in the output. In the main dialog box, click on OK to run the model.

The corresponding syntax is:

```
NOMREG
  newchd WITH cat age chl ecg smk hpt ch cc
  /CRITERIA = CIN(95) DELTA(0) MXITER(100) MXSTEP(5) LCONVERGE(0)
    PCONVERGE
  (1.0E-6) SINGULAR(1.0E-8)
  /MODEL
  /INTERCEPT = INCLUDE
  /PRINT = COVB PARAMETER SUMMARY LRT.
```

Note that the recoded CHD variable NEWCHD is used in the model statement. The NEWCHD value of zero corresponds to the CHD value of one.

The output is omitted.

The GENLIN procedure can be used to run GLM and GEE models, including unconditional logistic regression, which is a special case of GLM. To run the GENLIN procedure, select Analyze → Generalized Linear Models → Generalized Linear Models from the drop-down menus to reach a dialog box called "Type of Model." Click on Binary logistic under the heading called "Binary Response or Event/Trials Data." Next select a new dialogue box called "Response" and select the variable CHD in the Dependent Variable box. Click on the box called "Reference Category" and select First (lowest value) and click on Continue. Select a new dialogue box called "Predictors" and enter the covariates in the Covariates box. Select a new dialogue box called "Model" and enter the same covariates in the Model box. Click OK to run the model. The corresponding syntax follows (output omitted):

```
GENLIN chd (REFERENCE=FIRST) WITH age cat chl dbp ecg sbp smk hpt cc ch
  /MODEL age cat chl ecg smk hpt cc ch INTERCEPT=YES
DISTRIBUTION=BINOMIAL LINK=LOGIT
```

```
    /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=MODEL MAXITERATIONS=100
    MAXSTEPHALVING=5
    PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012 ANALYSISTYPE=3(WALD)
    CILEVEL=95
  CITYPE=WALD
    LIKELIHOOD=FULL
    /MISSING CLASSMISSING=EXCLUDE
    /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

Obtaining ROC Curves

The ROC procedure will produce ROC curves in SPSS. If we wish to use predicted probabilities from a logistic regression as cutpoints for an ROC curve, we must first run a logistic regression and save the predicted probabilities in our working dataset. Then we can use the ROC procedure. This will be demonstrated with the knee fracture dataset.

Open the dataset **kneefr.sav in** the Data Editor window. The corresponding command syntax is:

```
  GET
    FILE='C:\kneefr.sav'.
```

The outcome variable is FRACTURE indicating whether the subject actually had a knee fracture. The model follows:

$$\text{logit P(FRACTURE} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{AGECAT} + \beta_2\text{HEAD} + \beta_3\text{PATELLAR} + \beta_4\text{FLEX} + \beta_5\text{WEIGHT}$$

To run the LOGISTIC REGRESSION procedure, select Analyze → Regression → Binary Logistic from the drop-down menus to reach the dialog box to specify the logistic model. Select FRACTURE from the variable list and enter it into the Dependent Variable box, then select and enter the covariates AGECAT, HEAD, PATELLAR, FLEX, and WEIGHT into the Covariate(s) box. Click on SAVE to create a new variable in the knee fracture dataset. Check the box called "Probabilities" under the heading "Predicted Values." Select CONTINUE and then click on OK to run the model. A new variable called PRE_1 will appear in the working dataset containing each individual's predicted probability. These predicated probabilities are used to help generate the ROC curve.

The two key variables for producing an ROC curve using a logistic regression are the predicated probabilities (called PRE_1 in this example) and the observed dichotomous outcome variable (called FRACTURE in this example). To obtain an ROC curve select Analyze → ROC Curve, then select the variable Predicted probability (PRE_1) in the box called "Test Variable" and select the outcome variable FRACTURE in the box called "State Variable." Type the value 1 in the box called "Value of State Variable" since FRACTURE = 1 indicates a fracture event. Click on OK to obtain the ROC curve.

The corresponding syntax, to run the logistic regression, create the new variable PRE_1, and generate an ROC curve follows:

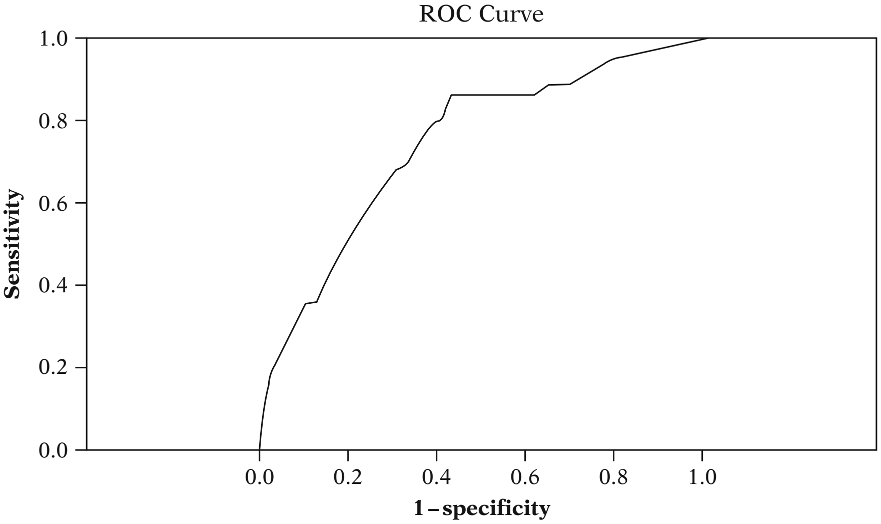
```
LOGISTIC REGRESSION VARIABLES fracture
  /METHOD=ENTER agecat head patellar flex weight
  /SAVE=PRED
  /CLASSPLOT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).

ROC PRE_1 BY fracture (1)
  /PLOT=CURVE
  /PRINT= COORDINATES
  /CRITERIA=CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING=EXCLUDE.
```

The output containing the parameter estimates of the logistic regression as well as the resultant ROC curve from the model follow:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(*B*)
Step 1[a]	agecat	.556	.399	1.938	1	.164	1.744
	head	.218	.376	.337	1	.562	1.244
	patellar	.627	.352	3.175	1	.075	1.872
	flex	.528	.374	1.988	1	.159	1.695
	weight	1.506	.409	13.532	1	.000	4.507
	Constant	−3.466	.412	70.837	1	.000	.031

[a]Variable(s) entered on step 1: agecat, head, patellar, flex, weight.



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s): Predicted probability

Area
.745

Conditional Logistic Regression

SPSS does not perform conditional logistic regression except in the *special case* in which there is only one case per stratum, with one or more controls. The SPSS survival analysis procedure COXREG can be used to obtain coefficient estimates equivalent to running a conditional logistic regression.

Recall that the MI dataset contains information on 39 cases diagnosed with myocardial infarction, each of which is matched with two controls. Thus, it meets the criterion of one case per stratum. A *time* variable must be created in the data, coded to indicate that all cases had the event at the same time, and all controls were censored at a later time. This variable has already been included in the SPSS dataset **mi.sav**. The variable has the value 1 for all cases and the value 2 for all controls.

The first step is to open the SPSS dataset, **mi.sav**, into the Data Editor window. The corresponding command syntax is:

```
GET
  FILE=‘C:\mi.sav’.
```

To run the equivalent of a conditional logistic regression analysis, select Analyze → Survival → Cox Regression from the drop-down menus to reach the dialog box to specify the model. Select SURVTIME from the variable list and enter it into the Time box. The Status box identifies the variable that indicates whether the subject had an event or was censored. For this dataset, select and enter MI into the Status box. The value of the variable that indicates that the event has occurred (i.e., that the subject is a case) must also be defined. This is done by clicking on the Define Event button and entering the value “1” in the new dialog box. Next, select and enter the covariates of interest (i.e., SMK, SBP, ECG) into the Covariate box. Finally, select and enter the variable which defines the strata in the Strata box. For the MI dataset, the variable is called MATCH. Click on OK to run the model.

The corresponding syntax, with the default specifications regarding the modeling process follows:

```
COXREG
  survtime /STATUS=mi(1) /STRATA=match
  /METHOD=ENTER smk sbp ecg
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```

The model statement contains the time variable (SURVTIME) followed by a backslash and the case status variable (MI) with the value for cases (1) in parentheses.

The output is omitted.

Polytomous Logistic Regression

Polytomous logistic regression is demonstrated with the cancer dataset using the NOMREG procedure described previously.

The outcome variable is SUBTYPE, a three-category outcome indicating whether the subject's histological subtype is Adenocarcinoma (coded 0), Adenosquamous (coded 1), or Other (coded 2). The model is restated as follows:

$$\ln\left[\frac{\mathrm{P}(\mathrm{SUBTYPE}=g|\mathbf{X})}{\mathrm{P}(\mathrm{SUBTYPE}=0|\mathbf{X})}\right]=\alpha_g+\ \beta_{g1}\mathrm{AGE}+\beta_{g2}\ \mathrm{ESTROGEN}+\beta_{g3}\ \mathrm{SMOKING},$$
$$\text{where } g=1,2$$

By default, the highest level of the outcome variable is the reference group in the NOMREG procedure. If we wish to make SUBTYPE = 0 (Adenocarcinoma) the reference group, as was done in the presentation in Chap. 12, the variable SUBTYPE must be recoded. The new variable created by the recode is called NEWTYPE and has already been included in the SPSS dataset **cancer.sav**. The command syntax used for the recoding was as follows:

```
RECODE
  subtype
  (2=0) (1=1) (0=2) INTO newtype.
EXECUTE.
```

To run the NOMREG procedure, select Analyze → Regression → Multinomial Logistic from the drop-down menus to reach the dialog box to specify the logistic model. Select NEWTYPE from the variable list and enter it into the Dependent Variable box, then select and enter the covariates (AGE, ESTROGEN, and SMOKING) into the Covariate(s) box. In the main dialog box, click on OK to run the model with the default settings.

The corresponding syntax is shown next, followed by the output generated by running the procedure.

```
NOMREG
  newtype WITH age estrogen smoking
  /CRITERIA = CIN(95) DELTA(0) MXITER(100) MXSTEP(5) LCONVERGE(0)
    PCONVERGE
  (1.0E-6) SINGULAR(1.0E-8)
  /MODEL
  /INTERCEPT = INCLUDE
  /PRINT = PARAMETER SUMMARY LRT.
```

Nominal Regression

Case Processing Summary

		N
NEWTYPE	.00	57
	1.00	45
	2.00	184
Valid		286
Missing		2
Total		288

Parameter Estimates

		B	Std. error	Wald	df	Sig.
NEWTYPE						
.00	Intercept	−1.203	.319	14.229	1	.000
	AGE	.282	.328	.741	1	.389
	ESTROGEN	−.107	.307	.122	1	.727
	SMOKING	−1.791	1.046	2.930	1	.087
1.00	Intercept	−1.882	.402	21.869	1	.000
	AGE	.987	.412	5.746	1	.017
	ESTROGEN	−.644	.344	3.513	1	.061
	SMOKING	.889	.525	2.867	1	.090

		Exp(*B*)	95% Confidence interval for Exp(*B*)	
NEWTYPE			Lower bound	Upper bound
.00	Intercept			
	AGE	1.326	.697	2.522
	ESTROGEN	.898	.492	1.639
	SMOKING	.167	2.144E-02	1.297
1.00	Intercept			
	AGE	2.683	1.197	6.014
	ESTROGEN	.525	.268	1.030
	SMOKING	2.434	.869	6.815

There are two parameter estimates for each independent variable and two intercepts. The estimates are grouped by comparison. The first set compares NEWTYPE = 0 to NEWTYPE = 2. The second comparison is for NEWTYPE = 1 to NEWTYPE = 2. With the original coding of the subtype variable, these are the comparisons of SUBTYPE = 2 to SUBTYPE = 0 and SUBTYPE = 1 to SUBTYPE = 0 respectively. The odds ratio for AGE = 1 vs. AGE = 0 comparing SUBTYPE = 2 vs. SUBTYPE = 0 is **exp**(0.282) = 1.33.

Ordinal Logistic Regression

Ordinal logistic regression is carried out in SPSS using the PLUM procedure. We again use the cancer dataset to demonstrate this model. For this analysis, the variable

GRADE is the response variable. GRADE has three levels, coded 0 for well differentiated, 1 for moderately differentiated, and 2 for poorly differentiated.

The model is stated as follows:

$$\ln\left[\frac{\mathrm{P(GRADE} \le g^{*}|\mathbf{X})}{\mathrm{P(GRADE} > g^{*}|\mathbf{X})}\right] = \alpha^{*}_{g^{*}} - \beta^{*}_{1}\mathrm{RACE} - \beta^{*}_{2}\ \mathrm{ESTROGEN\ for}\ g^{*} = 0, 1$$

Note that this is the alternative formulation of the ordinal model discussed in Chap. 13. In contrast to the formulation presented in the SAS section of the appendix, SPSS models the odds that the outcome is in a category less than or equal to category g^*. The other difference in the alternative formulation of the model is that there are negative signs before the beta coefficients. These two differences "cancel out" for the beta coefficients so that $\beta_i = \beta_i^{*}$ however, for the intercepts, $\alpha_g = -\alpha^{*}_{g^{*}}$, where α_g and β_i, respectively, denote the intercept and ith regression coefficient in the model run using SAS.

To perform an ordinal regression in SPSS, select Analyze → Regression → Ordinal from the drop-down menus to reach the dialog box to specify the logistic model. Select GRADE from the variable list and enter it into the Dependent Variable box, then select and enter the covariates (RACE and ESTROGEN) into the Covariate(s) box. Click on the Output button to request a "Test of Parallel Lines," which is a statistical test that SPSS provides that performs a similar function as the Score test of the proportional odds assumption in SAS. In the main dialog box, click on OK to run the model with the default settings.

The command syntax for the ordinal regression model is as follows:

```
PLUM grade WITH race estrogen
  /CRITERIA=CIN(95) DELTA(0) LCONVERGE(0) MXITER(100) MXSTEP(5)
  PCONVERGE(1.0E-6) SINGULAR(1.0E-8)
  /LINK=LOGIT
  /PRINT=FIT PARAMETER SUMMARY TPARALLEL.
```

The output generated by this code follows:

PLUM – Ordinal Regression

Test of Parallel Lines

Model	–2 Log likelihood	Chi-square	df	Sig.
Null hypothesis	34.743			
General	33.846	.897	2	.638

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

Parameter Estimates

	Estimate	Std. error	Wald	df	Sig.
Threshold [GRADE = .00]	−.511	.215	5.656	1	.017
[GRADE = 1.00]	1.274	.229	31.074	1	.000
Location RACE	.427	.272	2.463	1	.117
ESTROGEN	−.776	.249	9.696	1	.002

Link function: Logit

95% Confidence interval	
Lower bound	Upper bound
−.932	−8.981E-02
.826	1.722
−.106	.960
−1.265	−.288

A test of the parallel lines assumption is given in the table titled "Test of Parallel Lines." The null hypothesis is that the slope parameters are the same for the two different outcome comparisons (i.e., the proportional odds assumption). The results of the chi-square test statistic are not statistically significant ($p = 0.638$), suggesting that the assumption is tenable.

The parameter estimates and resulting odds ratios are given in the next table. As noted earlier, with the alternate formulation of the model, the parameter estimates for RACE and ESTROGEN match those of the SAS output, but the signs of the intercepts (labeled Threshold on the output) are reversed.

Modeling Correlated Dichotomous Data with GEE

The programming of a GEE model with the infant care dataset is demonstrated using the GENLIN procedure. The model is stated as follows:

$$\text{logit } P(\text{OUTCOME} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{BIRTHWGT} + \beta_2\text{GENDER} + \beta_3\text{DIARRHEA}$$

The dichotomous outcome variable (called OUTCOME) is derived from a weight-for-height standardized score based on the weight-for-height distribution of a standard population. The outcome is correlated since there are multiple measurements for each infant. The code and output are shown for this model assuming an AR1 correlation structure.

Open the dataset **infant.sav** in the Data Editor window. The corresponding command syntax is:

```
GET
  FILE='C:\infant.sav'.
```

To run GEE model using the GENLIN procedure, select Analyze → Generalized Linear Models → Generalized Estimating Equations from the drop-down menus to reach a dialog box called Repeated (the word Repeated is highlighted in the upper right corner). Select the variable IDNO in the box called “Subject variables” and select the variable MONTH in the box called “Within-subject variables”. Under the heading called “Covariance Matrix” there are two possible choices to click on: Robust Estimator or Model-based Estimator. Keep it on the default Robust Estimator. Below that is the heading called “Working Correlation Matrix.” To the right of the word Structure is the default choice of an independent working correlation structure. Click on the drop-down menu and you will see four other choices for the working correlation structure: AR(1), Exchangeable, M-dependent, and Unstructured. Click on AR(1).

Next select a new dialog box called “Type of Model.” Click on Binary logistic under the heading called “Binary Response or Event/Trials Data.” Select a new dialogue box called “Response” and select the variable OUTCOME in the Dependent Variable box. Click on the box called “Reference Category” and select First (lowest value) and click on Continue. Select a new dialogue box called “Predictors” and enter the variables BIRTHWGT, GENDER, and DIARRHEA in the Covariates box. Select a new dialogue box called “Model” and enter the same covariates in the Model box. Select a new dialogue box called “Statistics.” Under the heading Print, many of the output statistics are checked by default. Click on one that is not checked by default called “Working correlation matrix” (bottom left) and click OK to run the model. The corresponding syntax follows:

```
Generalized Estimating Equations.
GENLIN outcome (REFERENCE=FIRST) WITH birthwgt gender diarrhea
  /MODEL birthwgt gender diarrhea INTERCEPT=YES
DISTRIBUTION=BINOMIAL LINK=LOGIT
  /CRITERIA METHOD=FISHER(1) SCALE=1 MAXITERATIONS=100
  MAXSTEPHALVING=5
PCONVERGE=1E-006 (ABSOLUTE)
  SINGULAR=1E-012 ANALYSISTYPE=3(WALD) CILEVEL=95 LIKELIHOOD=FULL
  /REPEATED SUBJECT=idno WITHINSUBJECT=month SORT=YES CORRTYPE=AR(1)
ADJUSTCORR=YES COVB=ROBUST
  MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION WORKINGCORR.
```

Selected output follows:

Model Information

Dependent Variable	outcome[a]
Probability Distribution	Binomial
Link Function	Logit
Subject Effect 1	idno
Within-Subject Effect 1	month
Working Correlation Matrix Structure	AR(1)

[a]The procedure models 1.00 as the response, treating .00 as the reference category.

Parameter Estimates

			95% Wald confidence interval		Hypothesis test		
Parameter	*B*	Std. error	Lower	Upper	Wald chi-square	df	Sig.
(Intercept)	−1.398	1.1960	−3.742	.946	1.366	1	.243
birthwgt	.0005	.0003	−.001	.000	2.583	1	.108
gender	.002	.5546	−1.085	1.089	.000	1	.997
diarrhea	.221	.8558	−1.456	1.899	.067	1	.796
(Scale)	1						

Dependent Variable: outcome
Model: (Intercept), birthwgt, gender, diarrhea

Working Correlation Matrix[a]

	Measurement								
Measurement	[month = 1.00]	[month = 2.00]	[month = 3.00]	[month = 4.00]	[month = 5.00]	[month = 6.00]	[month = 7.00]	[month = 8.00]	[month = 9.00]
[month = 1.00]	1.000	.525	.276	.145	.076	.040	.021	.011	.006
[month = 2.00]	.525	1.000	.525	.276	.145	.076	.040	.021	.011
[month = 3.00]	.276	.525	1.000	.525	.276	.145	.076	.040	.021
[month = 4.00]	.145	.276	.525	1.000	.525	.276	.145	.076	.040
[month = 5.00]	.076	.145	.276	.525	1.000	.525	.276	.145	.076
[month = 6.00]	.040	.076	.145	.276	.525	1.000	.525	.276	.145
[month = 7.00]	.021	.040	.076	.145	.276	.525	1.000	.525	.276
[month = 8.00]	.011	.021	.040	.076	.145	.276	.525	1.000	.525
[month = 9.00]	.006	.011	.021	.040	.076	.145	.276	.525	1.000

Dependent Variable: outcome
Model: (Intercept), birthwgt, gender, diarrhea
[a] The AR(1) working correlation matrix structure is computed assuming the measurements are equally spaced for all subjects.

The output contains tables for GEE model information, GEE parameter estimates, and the working correlation matrix. The working correlation matrix is a 9 × 9 matrix with an AR1 correlation structure. The table containing the GEE parameter estimates uses the empirical standard errors by default. Model-based standard errors could also have been requested. The odds ratio estimate for DIARRHEA = 1 vs. DIARRHEA = 0 is **exp**(.221) = 1.247.

The SPSS section of this appendix is completed. Next, modeling with Stata software is illustrated.

STATA

Stata is a statistical software package that has become increasingly popular in recent years. Analyses are obtained by typing the appropriate statistical commands in the Stata Command window or in the Stata Do-file Editor window. The commands used to perform the statistical analyses in this appendix are listed below. These commands are case sensitive and lower case letters should be used. In the text, commands are given in bold font for readability.

1. **logit** – This command is used to run logistic regression.
2. **binreg** – This command can also be used to run logistic regression. The **binreg** command can also accommodate summarized binomial data in which each observation contains a count of the number of events and trials for a particular pattern of covariates.
3. **clogit** – This command is used to run conditional logistic regression.
4. **mlogit** – This command is used to run polytomous logistic regression.
5. **ologit** – This command is used to run ordinal logistic regression.
6. **xtset** – This command is used to define the cluster variable(s) for subsequent analyses of correlated data using Stata commands beginning with **xt**.
7. **xtgee** – This command is used to run GEE models.
8. **xtiogit** – This command can be used to run GEE logistic regression models.
9. **xtmelogit** – This command is used to run logistic mixed models.
10. **lrtest** – This command is used to perform likelihood ratio tests.
11. **lincom** – This command is used to calculate a linear combination of parameter estimates following a regression command.

Four windows will appear when Stata is opened. These windows are labeled Stata Command, Stata Results, Review, and Variables. As with SPSS, the user can click on File → Open to select a working dataset for analysis. Once a dataset is selected, the names of its variables appear in the Variables window. Commands are entered in the Stata Command window. The output generated by commands appears in the Results window after the enter key is pressed. The Review window preserves a history of all the commands executed during the Stata session. The commands in the Review window can be saved, copied, or edited as the user desires. Commands can also be run from the Review window by double-clicking on the command.

Alternatively, commands can be typed, or pasted into the Do-file Editor. The Do-file Editor window is activated by clicking on Window → Do-file Editor or by simply clicking on the Do-file Editor button on the Stata tool bar. Commands are executed from the Do-file Editor by clicking on Tools → Do. The advantage of running commands from the Do-file Editor is that commands need not be entered and executed one at a time as they do from the Stata Command window. The Do-file Editor serves a similar function as the Program Editor in SAS.

Unconditional Logistic Regression

Unconditional logistic regression is illustrated using the Evans County data. As discussed in the previous sections, the dichotomous outcome variable is CHD and the covariates are CAT, AGE, CHL, ECG, SMK, and HPT. Two interaction terms CH and CC, are also included. CH is the product CAT × HPT, while CC is the product CAT × CHL. The variables representing the interaction terms have already been included in the Stata dataset **evans.dta**.

The model is restated as follows:

$$\text{logit P(CHD} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{CAT} + \beta_2\text{AGE} + \beta_3\text{CHL} + \beta_4\text{ECG} + \beta_5\text{SMK} + \beta_6\text{HPT} + \beta_7\text{CH} + \beta_8\text{CC}$$

The first step is to activate the Evans dataset by clicking on File → Open and selecting the Stata dataset, **evans.dta**. The code to run the logistic regression is as follows:

```
logit chd cat age chl ecg smk hpt ch cc
```

Following the command **logit** comes the dependent variable followed by a list of the independent variables. Clicking on the variable names in the Variable Window pastes the variable names into the Command Window. For **logit** to run properly in Stata, the dependent variable must be coded zero for the nonevents (in this case, absence of coronary heart disease) and nonzero for the event. The output produced in the results window is as follows:

```
Iteration 0: log likelihood = −219.27915
Iteration 1: log likelihood = −184.11809
Iteration 2: log likelihood = −174.5489
Iteration 3: log likelihood = −173.64485
Iteration 4: log likelihood = −173.61484
Iteration 5: log likelihood = −173.61476


Logit estimates                                  Number of obs  =      609
                                                 LR chi2(8)     =    91.33
                                                 Prob > chi2    =   0.0000
Log likelihood = −173.61476                      Pseudo R2      =   0.2082

-------------------------------------------------------------------------
 chd   |    Coef.      Std. Err.      z     P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------
cat    |−12.68953     3.10465     −4.09   0.000   −18.77453     −6.604528
age    |   .0349634    .0161385    2.17   0.030     .0033327     .0665942
cht    |  −.005455     .0041837   −1.30   0.192    −.013655      .002745
ecg    |   .3671308    .3278033    1.12   0.263    −.275352     1.009614
smk    |   .7732135    .3272669    2.36   0.018     .1317822    1.414645
hpt    |  1.046649     .331635     3.16   0.002     .3966564    1.696642
ch     | −2.331785     .7426678   −3.14   0.002   −3.787387     −.8761829
cc     |   .0691698    .0143599    4.82   0.000     .0410249     .0973146
_cons  | −4.049738    1.255015    −3.23   0.001   −6.509521    −1.589955
-------------------------------------------------------------------------
```

The output indicates that it took five iterations for the log likelihood to converge at −173.61476. The iteration history appears at the top of the Stata output for all of the models illustrated in this appendix. However, we shall omit that portion of the output in subsequent examples. The table shows the regression coefficient estimates and standard error, the test statistic (*z*) and *p*-value for the Wald test, and 95% confidence intervals. The intercept, labeled "cons" (for constant), is given in the last row of the table. Also included in the output is a likelihood ratio test statistic (91.33) and corresponding p-value (0.0000) for a likelihood ratio test comparing the full model with 8 regression parameters to a reduced model containing only the intercept. The test statistic follows a chi-square distribution with 8 degrees of freedom under the null.

The **or** option for the **logit** command is used to obtain exponentiated coefficients rather than the coefficients themselves. In Stata, options appear in the command following a comma. The code follows:

```
    logit chd cat age chl ecg smk hpt ch cc, or
```

The **logistic** command without the **or** option produces identical output as the **logit** command does with the **or** option. The output follows:

```
Logit estimates                                       Number of obs   =      609
                                                      LR chi2 (8)     =    91.33
                                                      Prob > chi2     =   0.0000
Log likelihood = −173.61476                           Pseudo R2       =   0.2082
-----------------------------------------------------------------------------------
 chd  |  Odds Ratio    Std. Err.      z        P> |z|    [95% Conf. Interval]
------|----------------------------------------------------------------------------
cat   |  3.08e-06     9.57e-06     −4.09      0.000    7.02e-09     .0013542
age   |  1.035582      .0167127     2.17      0.030    1.003338    1.068862
chl   |   .9945599     .004161     −1.30      0.192     .9864378   1.002749
ecg   |  1.443587      .4732125     1.12      0.263     .7593048   2.74454
smk   |  2.166718      .709095      2.36      0.018    1.14086     4.115025
hpt   |  2.848091      .9445266     3.16      0.002    1.486845    5.455594
ch    |   .0971222     .0721295    −3.14      0.002     .0226547    .4163692
cc    |  1.071618      .0153883     4.82      0.000    1.041878    1.102207
-----------------------------------------------------------------------------------
```

The standard errors and 95% confidence intervals are those for the odds ratio estimates. As discussed in the SAS section of this appendix, care must be taken in the interpretation of these odds ratios with continuous predictor variables or interaction terms included in the model.

The **vce** command will produce a variance–covariance matrix of the parameter estimates. Use the **vce** command after running a regression. The code and output follow:

```
 vce  |
      |   cat        age         chl        ecg        smk        hpt       _cons
------|---------------------------------------------------------------------------
cat   |  .12389
age   | −.002003    .00023
chl   |  .000283  −2.3e-06    .000011
ecg   | −.027177   −.000105   .000041    .086222
smk   | −.006541    .000746   .00002     .007845    .093163
hpt   | −.032891   −.000026  −.000116  −.00888     .001708    .084574
_cons |  .042945   −.012314  −.002271  −.027447  −.117438  −.008195  1.30013
      |
```

The **lrtest** command can be used to perform likelihood ratio tests. For example, to perform a likelihood ratio test on the two interaction terms, CH and CC, in the preceding model, we can save the -2 log likelihood statistic of the full model in the computer's memory by using the command **estimates store** followed by a user defined name called **full** in this example:

```
estimates store full
```

Now the reduced model (without the interaction terms) can be run (output omitted):

```
logit chd cat age chl ecg smk hpt
```

After the reduced model is run, type the following command to obtain the results of the likelihood ratio test comparing the full model (with the interaction terms) to the reduced model:

```
Lrtest full
```

The resulting output follows:

```
Logit: likelihood-ratio test                chi2(2)    =  53.16
(Assumption . nested in full)              Prob > chi2 = 0.0000
```

The chi-square statistic with 2 degrees of freedom is 53.16, which is statistically significant as the *p*-value is close to zero.

The **lincom** command can be used to calculate a linear combination of parameters. As with the **vce** and **lrtest** commands, the **lincom** command is used directly after running a model. Recall, the code to run the full model with the two interaction terms is:

```
logit chd cat age chl ecg smk hpt ch cc, or
```

Now suppose we wish to estimate the odds ratio for CAT = 1 vs. CAT = 0 among those with HPT = 1 and CHOL = 220. This odds ratio is **exp** $(\beta_1 + 1\beta_6 + 220\beta_7)$, and can be estimated using the **lincom** command as follows:

```
lincom cat*1 + ch*1 + cc*220, or
```

The **or** option requests that the linear combination of parameter estimates be exponentiated. The output containing the odds ratio estimate using the **lincom** command follows:

```
----------------------------------------------------------------------
chd ¦ Odds Ratio   Std. Err.     z      P>|z|     [95% Conf. Interval]
----¦-----------------------------------------------------------------
(1) ¦  1.216568    .5808373    0.41    0.681    .4772429    3.101226
----¦-----------------------------------------------------------------
```

The Evans County dataset contains individual level data. In the SAS section of this appendix, we illustrated how to run a logistic regression on summarized binomial data in which each observation contained a count of the number of events and trials for a particular pattern of covariates. This can also be accomplished in Stata using the **binreg** command.

The summarized dataset, EVANS2, described in the SAS section contains eight observations and is small enough to be typed directly into the computer using the **input** command followed by a list of variables. The **clear** command clears the individual level Evans County dataset from the computer's memory and should be run before creating the new dataset since there are common variable names to the new and cleared dataset (CAT and ECG). After entering the **input** command, Stata will prompt you to enter each new observation until you type **end**.

The code to create the dataset is presented below. The newly defined five variables are described in the SAS section of this appendix.

```
clear

input cases total cat agegrp ecg

         cases      total      cat      agegrp      ecg
1.        17         274        0         0          0
2.        15         122        0         1          0
3.         7          59        0         0          1
4.         5          32        0         1          1
5.         1           8        1         0          0
6.         9          39        1         1          0
7.         3          17        1         0          1
8.        14          58        1         1          1
9.        end
```

The **list** command can be used to display the dataset in the Results Window and to check the accuracy of data entry.

The data is in binomial events/trials format in which the variable CASES represents the number of coronary heart disease cases and the variable TOTAL represents the number of subjects at risk in a particular stratum defined by the other three variables. The model is stated as follows:

$$\text{logit P(CHD} = 1) = \beta_0 + \beta_1\text{CAT} + \beta_2\text{AGEGRP} + \beta_3\text{ECG}$$

The code to run the logistic regression follows:

```
binreg cases cat age ecg, n(total)
```

The **n( )** option, with the variable TOTAL in parentheses, instructs Stata that TOTAL contains the number of trials for each stratum. The output is omitted.

Individual level data can also be summarized using frequency counts if the variables of interest are categorical variables. The dataset EVANS3, discussed in the SAS section, uses frequency weights to summarize the data. The variable COUNT contains the frequency of occurrences of each observation in the individual level data.

EVANS3 contains the same information as EVANS2 except that it has sixteen observations rather than eight. The difference is that with EVANS3, for each pattern of covariates there is an observation containing the frequency counts for CHD = 1 and another observation containing the frequency counts for CHD = 0. The code to create the data is:

```
clear

input chd cat agegrp ecg count

              chd       cat       agegrp       ecg        count
1.             1         0          0           0          17
2.             0         0          0           0          257
3.             1         0          1           0          15
4.             0         0          1           0          107
5.             1         0          0           1           7
6.             0         0          0           1          52
7.             1         0          1           1           5
8.             0         0          1           1          27
9.             1         1          0           0           1
10.            0         1          0           0           7
11.            1         1          1           0           9
12.            0         1          1           0          30
13.            1         1          0           1           3
14.            0         1          0           1          14
15.            1         1          1           1          14
16.            0         1          1           1          44
17.           end
```

The model is restated as follows:

$$\text{logit } P(\text{CHD} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{CAT} + \beta_2\text{AGEGRP} + \beta_3\text{ECG}$$

The code to run the logistic regression using the **logit** command with frequency weighted data is:

```
logit chd cat agegrp ecg [fweight = count]
```

The **[fweight =]** option, with the variable COUNT, instructs Stata that the variable COUNT contains the frequency counts. The **[fweight =]** option can also be used with the **binreg** command:

```
binreg chd cat agegrp ecg [fweight = count]
```

The output is omitted.

Obtaining ROC Curves

The knee fracture dataset will be used to illustrate how ROC curves are generated in Stata. Open the dataset **kneefr.dta**. The outcome variable is FRACTURE indicating whether the subject actually had a knee fracture: Five predictor variables will be used to obtain predicted probabilities from a logistic regression for each individual in the dataset. The model follows:

$$\begin{aligned}\text{logit P(FRACTURE} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{AGECAT} + \beta_2\text{HEAD} + \beta_3\text{PATELLAR}\\ + \beta_4\text{FLEX} + \beta_5\text{WEIGHT}\end{aligned}$$

The code to run this model is:

```
Logit fracture agecat head patellar flex weight, or
```

The output follows:

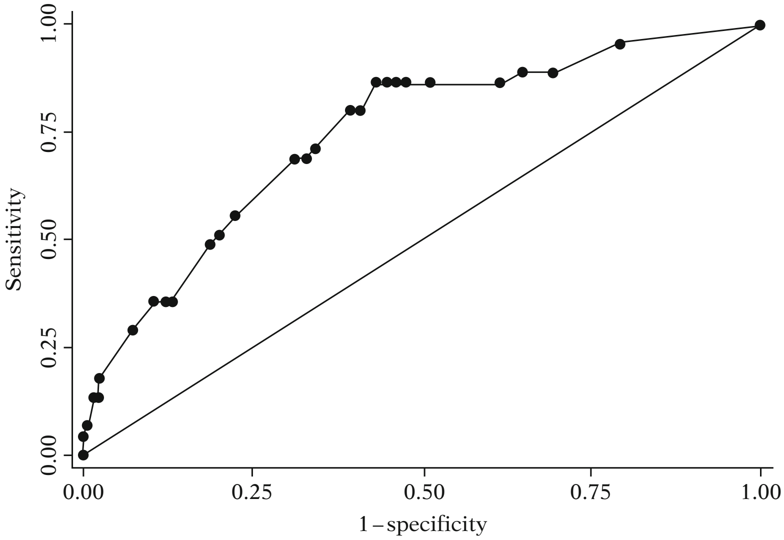
```
-----------------------------------------------------------------------------
fracture | Odds Ratio   Std. Err.      z     P>|z|     [95% Conf. Interval]
---------+-------------------------------------------------------------------
agecat   |  1.743647    .6964471    1.39   0.164     .7970246    3.814567
head     |  1.243907    .4678455    0.58   0.562     .595172     2.599758
patellar |  1.871685    .6584815    1.78   0.075     .9392253    3.729888
flex     |  1.695114    .6345218    1.41   0.159     .8139051    3.530401
weight   |  4.50681     1.844564    3.68   0.000    2.020628    10.05199
-----------------------------------------------------------------------------
```

Directly after running this model an ROC curve can be generated by using the **lroc** command. The code and output follows:

lroc

Logistic model for fracture

Number of observations = 348
Area under ROC curve = 0.7452



Area under ROC curve = 0.7452

The diagonal line across the plot serves as a reference line as to what would be expected if the predicted probabilities were uninformative. The area under this reference diagonal is 0.5. The area under the ROC curve is 0.745.

A slightly more complicated but more general approach for creating the same plot is to use the **roctab** command with the **graph** option. After running the logistic regression, the **predict** command can be used to create a new variable containing the predicted probabilities (named PROB in the code below). The **roctab** with the **graph** option is then used with the variable FRACTURE listed first as the true outcome and the newly created variable PROB listed next as the test variable. The code follows:

```
logit fracture agecat head patellar flex weight, or
predict prob
roctab fracture prob, graph
```

The **roctab** command can be used to create an ROC plot using any test variable against a true outcome variable. The test variable does not necessarily have to contain predicted probabilities from a logistic regression. In that sense, the **roctab** command is more general than the **lroc** command.

Conditional Logistic Regression

Conditional logistic regression is demonstrated with the MI dataset using the **clogit** command. The MI dataset contains information from a case-control study in which each case is matched with two controls. The model is stated as follows:

$$\text{logit P(CHD} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{SMK} + \beta_2\text{SPB} + \beta_3\text{ECG} + \sum_{i=1}^{38} \gamma_i V_i$$

$$V_i = \begin{cases} 1 \text{ if } i\text{th matched triplet} \\ 0 \text{ otherwise} \end{cases} \qquad i = 1, 2, \ldots, 38$$

Open the dataset **mi.dta.** The code to run the conditional logistic regression in Stata is:

```
clogit mi smk sbp ecg, strata (match)
```

The **strata()** option, with the variable MATCH in parentheses, identifies MATCH as the stratified variable (i.e., the matching factor). The output follows:

```
Conditional (fixed-effects) logistic regression  Number of obs   =       117
                                                 LR chi2 (3)     =     22.20
                                                 Prob > chi2     =    0.0001
Log likelihood = −31.745464                      Pseudo R2       =    0.2591
------------------------------------------------------------------------------
 mi  |    Coef.     Std. Err.     z      P>|z|      [95% Conf. Interval]
     |
-----|------------------------------------------------------------------------
smk  |   .7290581    .5612569    1.30    0.194    −.3709852    1.829101
sbp  |   .0456419    .0152469    2.99    0.003     .0157586     .0755251
ecg  |  1.599263     .8534134    1.87    0.061    −.0733967    3.271923
-----|------------------------------------------------------------------------
```

The **or** option can be used to obtain exponentiated regression parameter estimates. The code follows (output omitted):

```
    clogit mi smk sbp ecg, strata (match) or
```

Polytomous Logistic Regression

Polytomous logistic regression is demonstrated with the cancer dataset using the **mlogit** command.

The outcome variable is SUBTYPE, a three-category outcome indicating whether the subject's histological subtype is Adenocarcinoma (coded 0), Adenosquamous (coded 1), or Other (coded 2). The model is restated as follows:

$$\ln\left[\frac{\mathrm{P}(\mathrm{SUBTYPE}=g|\mathbf{X})}{\mathrm{P}(\mathrm{SUBTYPE}=0|\mathbf{X})}\right]=\alpha_g+\beta_{g1}\mathrm{AGE}+\beta_{g2}\mathrm{ESTROGEN}+\beta_{g3}\mathrm{SMOKING}$$

$$\text{where } g=1,2$$

Open the dataset **cancer.dta**. The code to run the polytomous logistic regression follows:

```
    mlogit subtype age estrogen smoking
```

Stata treats the outcome level that is coded zero as the reference group. The output follows:

```
Multinomial regression                              Number of obs   =      286
                                                    LR chi2(6)      =    18.22
                                                    Prob > chi2     =   0.0057
Log likelihood = −247.20254                         Pseudo R2       =   0.0355
----------------------------------------------------------------------------------
subtype  |     Coef.     Std. Err.     z     P>|z|      [95% Conf. Interval]
---------+------------------------------------------------------------------------
1        |
age      |   .9870592    .4117898    2.40   0.017     .179966      1.794152
estrogen |  −.6438991    .3435607   −1.87   0.061   −1.317266       .0294674
smoking  |   .8894643    .5253481    1.69   0.090    −.140199      1.919128
_cons    | −1.88218      .4024812   −4.68   0.000   −2.671029     −1.093331
---------+------------------------------------------------------------------------
2        |
age      |   .2822856    .3279659    0.86   0.389    −.3605158      .925087
estrogen |  −.1070862    .3067396   −0.35   0.727    −.7082847      .4941123
smoking  | −1.791312    1.046477    −1.71   0.087   −3.842369       .259746
_cons    | −1.203216     .3189758   −3.77   0.000   −1.828397      −.5780355
---------+------------------------------------------------------------------------
(Outcome subtype = = 0 is the comparison group)
```

Ordinal Logistic Regression

Ordinal logistic regression is demonstrated with the cancer dataset using the **ologit** command. For this analysis, the variable GRADE is the response variable. GRADE has three levels, coded 0 for well-differentiated, 1 for moderately differentiated, and 2 for poorly differentiated.

The model is stated as follows:

$$\ln\left[\frac{\mathrm{P}(\mathrm{GRADE} \le g^*|\mathbf{X})}{\mathrm{P}(\mathrm{GRADE} > g^*|\mathbf{X})}\right] = \alpha^*_{g^*} - \beta^*_1\mathrm{AGE} - \beta^*_2\mathrm{ESTROGEN}\ \text{for}\ g^* = 0, 1$$

This is the alternative formulation of the proportional odds model discussed in Chap. 13. In contrast to the formulation presented in the SAS section of the appendix, Stata, as does SPSS, models the odds that the outcome is in a category less than or equal to category g. The other difference in the alternative formulation of the model is that there are negative signs before the beta coefficients. These two differences “cancel out” for the beta coefficients so that $\beta_i = \beta_i^*$ however, for the intercepts, $\alpha_g = -\alpha^*_{g^*}$, where α_g and β_i, respectively, denote the intercept and ith regression coefficient in the model run using SAS.

The code to run the proportional odds model and output follows:

```
                        ologit grade race estrogen

Ordered logit estimates                          Number of obs   =      286
                                                 LR chi2 (2)     =    19.71
                                                 Prob > chi2     =   0.0001
Log likelihood = −287.60598                      Pseudo R2       =   0.0331
-----------------------------------------------------------------------------
grade     |   Coef.     Std. Err.      z       P>|z|     [95% Conf. Interval]
-----------------------------------------------------------------------------
race      |  .4269798   .2726439      1.57     0.117    −.1073926    .9613521
estrogen  | −.7763251   .2495253    −3.11     0.002   −1.265386    −.2872644
-----------------------------------------------------------------------------
_cut1     | −.5107035   .2134462              (Ancillary parameters)
_cut2     | 1.274351    .2272768
-----------------------------------------------------------------------------
```

Comparing this output to the corresponding output in SAS shows that the coefficient estimates are the same but the intercept estimates (labeled_cut1 and _cut2 in the Stata output) differ, as their signs are reversed due to the different formulations of the model.

Modeling Correlated Data with Dichotomous Outcomes

Stata has a series of commands beginning with the prefix **xt** that are designed for the analysis of longitudinal studies (sometimes called panel studies) with correlated outcomes. The first of **xt** commands that is typically used for analysis is the **xtset** command. This command defines the cluster variable and optionally a time variable indicating the time the observation was made within the cluster. We demonstrate some of the **xt** commands with the infant care dataset (**infant.dta**).

The variable in the infant dataset defining the cluster (infant) is IDNO. The variable defining the order of measurement within a cluster is MONTH. These variables can be set and then used in subsequent analyses with the **xtset** command. The code follows:

```
xtset idno month
```

Now when other **xt** commands are run using this dataset, the cluster and time variable do not have to be restated. The command **xtdescribe** can be typed to see descriptive measures of the cluster variable.

Next, a GEE model is demonstrated with the infant care dataset. GEE models can be executed with the **xtgee** command in Stata.

The model is stated as follows:

$$\text{logit } P(\text{OUTCOME} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{BIRTHWGT} + \beta_2\text{GENDER} + \beta_3\text{DIARRHEA}$$

The code to run this model with an AR1 correlation structure is:

```
xtgee outcome birthwgt gender diarrhea, family (binomial)
link(logit) corr(ar1) vce(robust)
```

Following the command **xtgee** is the dependent variable followed by a list of the independent variables. The **link()** and **family()** options define the link function and the distribution of the response. The **corr()** option allows the correlation structure to be specified. The **vce(robust)** option requests empirically based standard errors. The options **corr(ind), corr(exc), corr(sta4)**, and **corr(uns)**, can be used to request an independent, exchangeable, stationary 4-dependent, and an unstructured working correlation structure respectively.

The output using the AR1 correlation structure follows:

```
GEE population-averaged model                  Number of obs       =      1203
Group and time vars:              idno month   Number of groups    =       136
Link:                                logit     Obs per group: min  =         5
Family:                            binomial                   avg  =       8.8
Correlation:                         AR(1)                    max  =         9
                                               Wald chi2(3)        =      2.73
Scale parameter:                         1     Prob > chi2         =    0.4353

                                 (standard errors adjusted for clustering on idno)
----------------------------------------------------------------------------------
         |              Semi-robust
 outcome |    Coef.       Std. Err.       z     P > |z|     [95% Conf. Interval]
---------+------------------------------------------------------------------------
birthwgt |  −.0004942     .0003086     −1.60   0.109    −.0010991      .0001107
gender   |   .0023805     .5566551      0.00   0.997    −1.088643     1.093404
diarrhea |   .2216398     .8587982      0.26   0.796    −1.461574     1.904853
_cons    | −1.397792     1.200408      −1.16   0.244    −3.750549      .9549655
----------------------------------------------------------------------------------
```

The output does not match the SAS output exactly due to different estimation techniques but the results are very similar. If odds ratios are desired rather than the regression coefficients, then the **eform** option can be used to exponentiate the regression parameter estimates. The code and output using the **eform** option follow:

```
    xtgee outcome birthwgt gender diarrhea, family (binomial)
    link (logit) corr (ar1) robust eform
```

```
GEE population-averaged model                 Number of obs          =     1203
Group and time vars:               idno month  Number of groups       =      136
Link:                                logit     Obs per group: min     =        5
Family:                            binomial                   avg     =      8.8
Correlation:                         AR(1)                    max     =        9
                                              Wald chi2(3)           =     2.73
Scale parameter:                       1      Prob > chi2            =   0.4353

                                (standard errors adjusted for clustering on idno)
---------+--------------------------------------------------------------
         |               Semi-robust
outcome  | Odds Ratio     Std. Err.        z      P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------
birthwgt |   .9995059      .0003085     −1.60   0.109    .9989015    1.000111
gender   |  1.002383       .5579818      0.00   0.997    .3366729    2.984417
diarrhea |  1.248122      1.071885       0.26   0.796    .2318711    6.718423
---------+--------------------------------------------------------------
```

The **xtcorr** command can be used after running the GEE model to output the working correlation matrix. The code and output follow:

```
    xtcorr
```

```
Estimated within-idno correlation matrix R:

        c1      c2      c3      c4      c5      c6      c7      c8      c9
r1  1.0000
r2  0.5252  1.0000
r3  0.2758  0.5252  1.0000
r4  0.1448  0.2758  0.5252  1.0000
r5  0.0761  0.1448  0.2758  0.5252  1.0000
r6  0.0399  0.0761  0.1448  0.2758  0.5252  1.0000
r7  0.0210  0.0399  0.0761  0.1448  0.2758  0.5252  1.0000
r8  0.0110  0.0210  0.0399  0.0761  0.1448  0.2758  0.5252  1.0000
r9  0.0058  0.0110  0.0210  0.0399  0.0761  0.1448  0.2758  0.5252  1.0000
```

The same results could have been obtained with the **xtlogit** command. The **xtlogit** command is designed specifically for logistic regression with clustered data. The following code runs the same GEE model as shown above with the **xtlogit** command:

```
    xtlogit outcome birthwgt gender diarrhea, pa corr(ar1) vce(robust) or
```

The code is similar to that which was used with the **xtgee** command except that the **link()** and **family()** options found with the **xtgee** command are unnecessary with the **xtlogit** command as it is understood that a logistic regression is requested. The **pa** option (for population averaged) in **xtlogit** requests that a GEE model be run. The **or** option (for odds ratios) requests that the parameter estimates be exponentiated.

We have seen that the **xtlogit** command with the **pa** option requests a GEE model. If instead the **fe** option is used (for fixed effects) then the **xtlogit** command requests a conditional logistic regression to be run. Next, we consider a model with fixed effects (dummy variables) for the cluster variable. The model is stated as follows:

$$\text{logit } P(\text{OUTCOME} = 1|\mathbf{X}) = \beta_0 + \beta_1\text{BIRTHWGT} + \beta_2\text{GENDER} + \beta_3\text{DIARRHEA} + \sum_{i=1}^{135} \gamma_i V_i$$

$$V_i = \begin{cases} 1 & \text{if } i\text{th matched triplet} \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, 2, \ldots, 135$$

The indicator variables in this model are fixed effects. The code to run this model is shown two ways: first with the **xtlogit** command and then, for comparison, with the **clogit** command that was previously shown in the section on conditional logistic regression.

```
    xtlogit outcome birthwgt gender diarrhea, fe or
    clogit outcome birthwgt gender diarrhea, strata(idno) or
```

Both commands yield the same output. We show the output from the **xtlogit** command with the **fe** option:

```
note: 115 groups (1019 obs) dropped because of all positive or all negative
      outcomes.
note: birthwgt omitted because of no within-group variance.
note: gender omitted because of no within-group variance.

Iteration 0:   log likelihood = −64.410959
Iteration 1:   log likelihood = −64.409171
Iteration 2:   log likelihood = −64.409171

Conditional fixed-effects logistic regression  Number of obs      =      184
Group variable: idno                           Number of groups   =       21
                                               Obs per group: min =        7
                                                              avg =      8.8
                                                              max =        9
                                               LR chi2(1)         =     1.26
Log likelihood = −64.409171                    Prob > chi2        =   0.2615
---------------------------------------------------------------------------
 outcome  ¦     OR       Std. Err.     z     P>|z|     [95% Conf. Interval]
---------¦-----------------------------------------------------------------
diarrhea  ¦  2.069587    1.320018    1.14    0.254     .5928878    7.224282
---------¦-----------------------------------------------------------------
```

What is noteworthy with this output is that there are no parameter estimates for the variables BIRTHWGT and GENDER because the values of these variables do not vary within a cluster (infant). This is a key feature of conditional logistic regression. An intercept is not estimated for the same reason. As noted at the top of the output, even the variable of interest DIARRHEA, did not vary within most (115) of the infants. The data from only 21 infants were used for this analysis.

A model with a random effect for each infant can be run two ways in Stata: one way is with the **xtlogit** command and the **re** option and the other way with the **xtmelogit** command. A model with a random effect (a random intercept) for infant is stated as follows:

$$\begin{aligned}\text{logit } P(\text{OUTCOME} = 1|\mathbf{X}) = (\beta_0 + b_{0i}) + \beta_1\text{BIRTHWGT} + \beta_2\text{GENDER}\\ + \beta_3\text{DIARRHEA},\end{aligned}$$

where b_{0i} represents the random effect for subject i and is normally distributed with mean = 0 and variance = $\sigma_s{}^2$ (i.e., $b_{0i} \sim N(0, \sigma_s{}^2)$)

The code and output for the model using the **xtlogit** command with the **re** option (the default option) follow:

```
xtlogit outcome birthwgt gender diarrhea, re or
```

The **or** option requests the parameter estimates be exponentiated. The output follows:

```
Random−effects logistic regression     Number of obs          =     1203
Group variable: idno                   Number of groups       =      136
Random effects u_i ~ Gaussian          Obs per group: min     =        5
                                                      avg     =      8.8
                                                      max     =        9
                                       Wald chi2(3)           =     3.69
Log likelihood = −164.50654            Prob > chi2            =   0.2973
---------+-------------------------------------------------------------
 outcome |      OR       Std. Err.      z      P>|z|    [95% Conf. Interval]
---------+-------------------------------------------------------------
birthwgt |  .999352     .0005498    −1.18   0.239    .998275    1.00043
gender   | 1.59722     1.26687       0.59   0.555    .3374548   7.559861
diarrhea | 2.638831    1.753562      1.46   0.144    .717413    9.706306
---------+-------------------------------------------------------------
/lnsig2u | 2.360601     .276463                     1.818743    2.902458
---------+-------------------------------------------------------------
sigma_u  | 3.255352     .4499922                    2.482762    4.268358
rho      |  .7631004    .0499785                     .6520122    .8470451
---------+-------------------------------------------------------------
Likelihood−ratio test of rho = 0: chibar2(01) = 161.04 Prob > =  chibar2 =
0.000
```

The estimated odds ratio for DIARRHEA is 2.64, but it is not statistically significant as the standard error is large at 1.75 (p-value = 0.144). The standard deviation of the random effect, σ_s^2, is estimated at 3.255 to the right of the heading "sigma_u." The estimated variance of the random effect can be found by squaring the standard deviation ($3.255^2 = 10.6$). The natural log of the variance estimate is also given in the output at 2.36, to the right of the heading "/lnsig2u." The estimated value for "rho" is 0.763, which is the proportion of the total variance contributed by the subject specific variance component. The likelihood ratio test for the random effect is highly significant (chi-square = 61.04) at the bottom of the output.

An alternative command for running this model is the **xtmelogit** command. The **xtmelogit** command is designed to run mixed logistic models. That is, a mixing of fixed and random effects. The code to run the random intercept model with the **xtmelogit** command follows:

```
    xtmelogit outcome birthwgt gender diarrhea || idno: , or intpoints(3)
```

The symbol || separates the fixed effects from the subject specific random effect. The **or** option requests the parameter estimates for the fixed effects be exponentiated. The **intpoints(3)** option sets the number of quadrature points to 3 for the numerical estimation rather than the default 7. It is generally recommended to have more rather than fewer quadrature points for estimation. However, this model did not numerically estimate without this option to lower the default number, which likely indicates a problem with model stability. The output follows:

```
Mixed-effects logistic regression              Number of obs      =    1203
Group variable: idno                           Number of groups   =     136
                                               Obs per group: min =       5
                                                              avg =     8.8
                                                              max =       9
Integration points = 3                        Wald chi2(3)        =    4.26
Log likelihood = −170.33099                    Prob > chi2        =  0.2352
----------------------------------------------------------------------------
outcome  | Odds Ratio   Std. Err.     z     P>|z|    [95% Conf. Interval].
---------|------------------------------------------------------------------
birthwgt |  .9992259    .0005833   −1.33   0.185    .9980833     1.00037
gender   | 1.706531    1.450536     0.63   0.529    .3225525     9.028756
diarrhea | 2.763684    1.856623     1.51   0.130    .7407258    10.31144
----------------------------------------------------------------------------
Random-effects Parameters | Estimate   Std. Err.   [95% Conf. Interval]
--------------------------|-------------------------------------------------
idno: Identity            |
                 sd(_cons)| 2.732656   .5458486   1.847385   4.042153
----------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) = 149.39 Prob> = chibar2 = 0.0000
```

The estimated odds ratio for DIARRHEA is 2.76, but is not statistically significant (p-value = 0.130). The standard deviation of the random effect, σ_s^2, is estimated at 2.73 to the right of the heading "sd (_cons)." The results are similar but not identical to that obtained with the **xtlogit** command as the method of estimation is somewhat different.

The next model includes a random slope for the variable DIARRHEA in addition to the random intercept. The model is stated as follows:

$$\text{logit } P(\text{OUTCOME} = 1|\mathbf{X}) = (\beta_0 + b_{0\text{i}}) + \beta_1\text{BIRTHWGT} + \beta_2\text{GENDER} + (\beta_3 + b_{3i})\text{DIARRHEA},$$

where b_{0i} represents the random intercept for subject i and where b_{3i} represents a random slope with the variable DIARRHEA for subject $ib_{0i} \sim \mathbf{N}(0, \sigma_s{}^2)$ and $b_{3i} \sim \mathbf{N}(0, \sigma_0^2)$

This type of model cannot be run with the **xtlogit** command but can with the **xtmelogit** command. The code for running this model with two random effects follows:

```
xtmelogit outcome birthwgt gender diarrhea || idno: diarrhea,
     intpoints (3) covariance (uns) or
```

The code is similar to what was shown with the previous random intercept model except that DIARRHEA is added as a random effect (after idno:) as well as a fixed effect. Since this model contains more than one random effect we must consider the covariance between the random effects. The **covariance(uns)** requests an unstructured covariance structure for the random effects. The default **independent** covariance structure or the **exchangeable** covariance structure could also be requested for the random effects. The output follows:

```
Mixed-effects logistic regression                 Number of obs      =     1203
Group variable: idno                              Number of groups   =      136
                                                  Obs per group: min =        5
                                                                 avg =      8.8
                                                                 max =        9
Integration points = 3                            Wald chi2 (3)      =     2.10
Log likelihood = -167.91199                       Prob > chi2        =   0.5511
---------+----------------------------------------------------------------
 outcome | Odds Ratio   Std. Err.     z     P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------
birthwgt |   .9993232    .0006149   -1.10   0.271    .9981188    1.000529
gender   |  1.829316    1.610209     0.69   0.493    .3258667   10.26922
diarrhea |  9.50e-06     .0001605   -0.68   0.494   3.92e-20    2.30e+09
---------+----------------------------------------------------------------

---------------------------+----------------------------------------------
Random-effects Parameters  |  Estimate    Std. Err.     [95% Conf. Interval]
---------------------------+----------------------------------------------
idno: Unstructured         |
            sd (diarrhea)  | 10.5301     12.99674   .9372181   118.3109
                sd(_cons)  |  2.779649    .5656848  1.865359     4.142071
    corr(diarrhea, _cons)  |   .611977    .1909544    .11323     .8643853
---------------------------+----------------------------------------------
LR test vs. logistic regression: chi2(3) = 154.23 Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

This model is numerically very unstable and the confidence interval for DIARRHEA is basically estimated from 0 to infinity, which is not useful. There are three random effect parameters estimated: the standard deviation of the random slope for DIARRHEA, the random intercept, and the correlation between them. A likelihood ratio test of the three random effect parameters is given at the bottom of the output.

The **xtmelogit** command does not allow autocorrelation of the residuals to be modeled along with the random effects but rather assumes that the residuals have an independent correlation structure. However, the **xtmelogit** command does provide estimates for nested random effects. As a hypothetical example, suppose 30 daycare centers were randomly sampled and within each daycare center 10 infants were sampled yielding 300 infants in all (30 × 10). Also, each infant has monthly measurements over a 9-month period. In this setting, we can consider three types of independent variables: (1) a variable like DIARRHEA whose status may vary within an infant from month-to-month, (2) a variable like GENDER which is fixed at the infant level (does not vary month-to-month), and (3) a variable that is fixed at the daycare level such as the size of the daycare center. Here we have a cluster of daycare centers and nested within each daycare center is a cluster of infants. In the infant dataset, the variable identifying each infant is called IDNO. Suppose the variable identifying the daycare center was called DAYCARE (this variable does not actually exist in the infant dataset). Consider a model with a random intercept for each infant as well as a random intercept for each daycare center. We continue to use BIRTHWEIGHT, GENDER, and DIARRHEA as fixed effects. The code to run such a model using the **xtmelogit** command is:

```
xtmelogit outcome birthwgt gender diarrhea || daycare: || idno:
```

This model contains a random intercept at the daycare level and a random intercept at the infant level. The symbol "||" separating the random effects indicates that the random effect for infant (IDNO) is nested within DAYCARE. Random slope parameters could be listed after the code "`|| daycare:`" if they are to vary by daycare or listed after the code "`|| idno:`" if they are to vary by infant.

This completes our discussion on the use of SAS, SPSS, and STATA to run different types of logistic models. An important issue for all three of the packages discussed is that the user must be aware of how the outcome event is modeled for a given package and given type of logistic model. If the parameter estimates are the negative of what is expected, this could be an indication that the outcome value is not correctly specified for the given package and/or procedure.

All three statistical software packages presented have built-in Help functions which provide further details about the capabilities of the programs. The web-based sites of the individual companies are another source of information about the packages: http://www.sas.com/ for SAS, http://www.spss.com/ for SPSS, and http://www.stata.com/ for Stata.

Test Answers

Chapter 1

True-False Questions:

1. F: any type of independent variable is allowed
2. F: dependent variable must be dichotomous
3. T
4. F: S-shaped
5. T
6. T
7. F: cannot estimate risk using case-control study
8. T
9. F: constant term can be estimated in follow-up study
10. T
11. T
12. F: logit gives log odds, not log odds ratio
13. T
14. F: β_i controls for other variables in the model
15. T
16. F: multiplicative
17. F: $\exp(\beta)$ where β is coefficient of exposure
18. F: OR for effect of SMK is exponential of coefficient of SMK
19. F: OR requires formula involving interaction terms
20. F: OR requires formula that considers coding different from (0, 1)
21. e. $\exp(\beta)$ is not appropriate for *any X*.
22. $$\mathrm{P}(\mathbf{X}) = 1/(1 + \exp\{-[\alpha + \beta_1(\mathrm{AGE}) + \beta_2(\mathrm{SMK}) + \beta_3(\mathrm{SEX}) + \beta_4(\mathrm{CHOL}) + \beta_5(\mathrm{OCC})]\}).$$
23. $$\hat{\mathrm{P}}(\mathbf{X}) = 1/(1 + \exp\{-[-4.32 + 0.0274(\mathrm{AGE}) + 0.5859(\mathrm{SMK}) + 1.1523(\mathrm{SEX}) + 0.0087(\mathrm{CHOL}) - 0.5309(\mathrm{OCC})]\}).$$
24. $$\mathrm{logit}\ \mathrm{P}(\mathbf{X}) = -4.32 + 0.0274(\mathrm{AGE}) + 0.5859(\mathrm{SMK}) + 1.1523(\mathrm{SEX}) + 0.0087(\mathrm{CHOL}) - 0.5309(\mathrm{OCC}).$$

25. For a 40-year-old male smoker with CHOL = 200 and OCC = 1, we have

$$\mathbf{X} = (\text{AGE} = 40, \text{SMK} = 1, \text{SEX} = 1, \text{CHOL} = 200, \text{OCC} = 1),$$

assuming that SMK and SEX are coded as SMK = 1 if smoke, 0 otherwise, and SEX = 1 if male, 0 if female, and

$$\begin{aligned}\hat{P}(\mathbf{X}) &= 1/(1 + \exp\{-[-4.32 + 0.0274(40) + 0.5859(1) \\ &\quad + 1.1523(1) + 0.0087(200) - 0.5309(1)]\}) \\ &= 1/\{1 + \exp[-(-0.2767)]\} \\ &= 1/(1 + 1.319) \\ &= 0.431.\end{aligned}$$

26. For a 40-year-old male *non*smoker with CHOL = 200 and OCC = 1, **X** = (AGE = 40, SMK = 0, SEX = 1, CHOL = 200, OCC = 1)

and

$$\begin{aligned}\hat{P}(\mathbf{X}) &= 1/(1 + \exp\{-[-4.32 + 0.0274(40) + 0.5859(0) \\ &\quad + 1.1523(1) + 0.0087(200) - 0.5309(1)]\}) \\ &= 1/\{1 + \exp[-(-0.8626)]\} \\ &= 1/(1 + 2.369) \\ &= 0.297\end{aligned}$$

27. The RR is estimated as follows:

$$\begin{aligned}&\frac{\hat{P}(\text{AGE}=40, \text{SMK}=1, \text{SEX}=1, \text{CHOL}=200, \text{OCC}=1)}{\hat{P}(\text{AGE}=40, \text{SMK}=0, \text{SEX}=1, \text{CHOL}=200, \text{OCC}=1)} \\ &= 0.431/0.297 \\ &= 1.45\end{aligned}$$

This estimate can be interpreted to say smokers have 1.45 times as high a risk for getting hypertension as nonsmokers, controlling for age, sex, cholesterol level, and occupation.

28. If the study design had been case-control or cross-sectional, the risk ratio computation of Question 27 would be inappropriate because a risk or risk ratio cannot be directly estimated by using a logistic model unless the study design is follow-up. More specifically, the constant term α cannot be estimated from case-control or cross-sectional studies.

29. $\widehat{\text{OR}}$ (SMK controlling for AGE, SEX, CHOL, OCC)

$= e^{\hat{\beta}}$ where $\hat{\beta} = 0.5859$ is the coefficient of SMK in the fitted model

$= \exp(0.5859)$

$= 1.80$

This estimate indicates that smokers have 1.8 times as high a risk for getting hypertension as nonsmokers, controlling for age, sex, cholesterol, and occupation.

30. The rare disease assumption.
31. The odds ratio is a legitimate measure of association and could be used even if the risk ratio cannot be estimated.
32. $\widehat{\text{OR}}$ (OCC controlling for AGE, SEX, SMK, CHOL)

$$= e^{\hat{\beta}}, \text{where } \hat{\beta} = -0.5309 \text{ is the coefficient of OCC in the fitted model}$$
$$= \exp(-0.5309)$$
$$= 0.5881 = 1/1.70.$$

This estimate is less than 1 and thus indicates that unemployed persons (OCC = 0) are 1.70 times more likely to develop hypertension than are employed persons (OCC = 1).

33. Characteristic 1: the model contains only main effect variables
Characteristic 2: OCC is a (0, 1) variable.
34. The formula $\exp(\beta_i)$ is inappropriate for estimating the effect of AGE controlling for the other four variables because AGE is being treated as a continuous variable in the model, whereas the formula is appropriate for (0, 1) variables only.

Chapter 2

True-False Questions:

1. F: OR = $\exp(\psi)$
2. F: risk = $1/[1 + \exp(-\alpha)]$
3. T
4. T
5. T
6. T
7. T
8. F: OR = $\exp(\beta + 5\delta)$
9. The model in logit form is given as follows:

$$\begin{aligned}\text{logit } P(\mathbf{X}) = \alpha &+ \beta\text{CON} + \gamma_1\text{PAR} + \gamma_2\text{NP} + \gamma_3\text{ASCM} \\ &+ \delta_1\text{CON} \times \text{PAR} + \delta_2\text{CON} \times \text{NP} \\ &+ \delta_3\text{CON} \times \text{ASCM}.\end{aligned}$$

10. The odds ratio expression is given by

$$\exp(\beta + \delta_1\text{PAR} + \delta_2\text{NP} + \delta_3\text{ASCM}).$$

Chapter 3

1. a. $\text{ROR} = \exp(\beta)$
   b. $\text{ROR} = \exp(5\beta)$
   c. $\text{ROR} = \exp(2\beta)$
   d. All three estimated odds ratios should have the same value.
   e. The β in part b is one-fifth the β in part a; the β in part c is one-half the β in part a.
2. a. $\text{ROR} = \exp(\beta + \delta_1\text{AGE} + \delta_2\text{CHL})$
   b. $\text{ROR} = \exp(5\beta + 5\delta_1\text{AGE} + 5\delta_2\text{CHL})$
   c. $\text{ROR} = \exp(2\beta + 2\delta_1\text{AGE} + 2\delta_2\text{CHL})$
   d. For a given specification of AGE and CHL, all three estimated odds ratios should have the same value.
   e. The β in part b is one-fifth the β in part a; the β in part c is one-half the β in part a. The same relationships hold for the three δ_1s and the three δ_2s.
3. a. $\text{ROR} = \exp(5\beta + 5\delta_1\text{AGE} + 5\delta_2\text{SEX})$
   b. $\text{ROR} = \exp(\beta + \delta_1\text{AGE} + \delta_2\text{SEX})$
   c. $\text{ROR} = \exp(\beta + \delta_1\text{AGE} + \delta_2\text{SEX})$
   d. For a given specification of AGE and SEX, the odds ratios in parts b and c should have the same value.
4. a. $\text{logit P}(\mathbf{X}) = \alpha + \beta_1 S_1 + \beta_2 S_2 + \gamma_1\text{AGE} + \gamma_2\text{SEX}$, where S_1 and S_2 are dummy variables which distinguish between the three SSU groupings, e.g., $S_1 = 1$ if low, 0 otherwise and $S_2 = 1$ if medium, 0 otherwise.
   b. Using the above dummy variables, the odds ratio is given by $\text{ROR} = \exp(-\beta_1)$, where $\mathbf{X}^* = (0, 0, \text{AGE}, \text{SEX})$ and $\mathbf{X}^{**} = (1, 0, \text{AGE}, \text{SEX})$.
   c. $$\begin{aligned}\text{logit }\mathbf{P}(\mathbf{X}) = \alpha + \beta_1 S_1 + \beta_2 S_2 + \gamma_1\text{AGE} + \gamma_2\text{SEX} \\ +\delta_1(S_1 \times \text{AGE}) + \delta_2(S_1 \times \text{SEX}) \\ + \delta_3(S_2 \times \text{AGE}) + \delta_4(S_2 \times \text{SEX})\end{aligned}$$
   d. $\text{ROR} = \exp(-\beta_1 - \delta_1\text{AGE} - \delta_2\text{SEX})$
5. a. $\text{ROR} = \exp(10\beta_3)$
   b. $\text{ROR} = \exp(195\beta_1 + 10\beta_3)$
6. a. $\text{ROR} = \exp(10\beta_3 + 10\delta_{31}\text{AGE} + 10\delta_{32}\text{RACE})$
   b. $$\begin{aligned}\text{ROR} = \exp(195\beta_1 + 10\beta_3 + 195\delta_{11}\text{AGE} \\ +195\delta_{12}\text{RACE} + 10\delta_{31}\text{AGE} + 10\delta_{32}\text{RACE})\end{aligned}$$

Chapter 4

True-False Questions:

1. T
2. T
3. F: unconditional
4. T
5. F: the model contains a large number of parameters
6. T

7. T
8. F: α is not estimated in conditional ML programs
9. T
10. T
11. F: the variance–covariance matrix gives variances and covariances for regression coefficients, not variables.
12. T
13. Because matching has been used, the method of estimation should be *conditional* ML estimation.
14. The variables AGE and SOCIOECONOMIC STATUS do not appear in the printout because these variables have been matched on, and the corresponding parameters are nuisance parameters that are not estimated using a conditional ML program.
15. The OR is computed as e to the power 0.39447, which equals 1.48. This is the odds ratio for the effect of pill use adjusted for the four other variables in the model. This odds ratio says that pill users are 1.48 times as likely as nonusers to get cervical cancer after adjusting for the four other variables.
16. The OR given by e to -0.24411, which is 0.783, is the odds ratio for the effect of vitamin C use adjusted for the effects of the other four variables in the model. This odds ratio says that vitamin C is some-what protective for developing cervical cancer. In particular, since 1/0.78 equals 1.28, this OR says that vitamin C *nonusers* are 1.28 times more likely to develop cervical cancer than *users*, adjusted for the other variables.
17. Alternative null hypotheses:
    1. The OR for the effect of VITC adjusted for the other four variables equals 1.
    2. The coefficient of the VITC variable in the fitted logistic model equals 0.
18. The 95% CI for the effect of VITC adjusted for the other four variables is given by the limits 0.5924 and 1.0359.
19. The Z statistic is given by $Z = -0.24411/0.14254 = -1.71$
20. The value of MAX LOGLIKELIHOOD is the logarithm of the maximized likelihood obtained for the fitted logistic model. This value is used as part of a likelihood ratio test statistic involving this model.

Chapter 5

1. Conditional ML estimation is the appropriate method of estimation because the study involves matching.
2. Age and socioeconomic status are missing from the printout because they are matching variables and have been accounted for in the model by nuisance parameters which are not estimated by the conditional estimation method.
3. H_0: $\beta_{\text{SMK}} = 0$ in the no interaction model (Model I), or alternatively, H_0: OR $= 1$, where OR denotes the odds ratio for the effect of SMK on cervical cancer status, adjusted for the other variables (NS and AS) in model I; test statistic: Wald statistic $Z = \frac{\hat{\beta}_{\text{SMK}}}{S_{\hat{\beta}_{\text{SMK}}}}$, which is approximately normal (0, 1) under H_0, or alternatively, Z^2 is approximately chi square with one degree of freedom under H_0; test computation: $Z = \frac{1.4361}{0.3167} = 4.53$; alternatively, $Z^2 = 20.56$; the one-tailed P-value is $0.0000/2 = 0.0000$, which is highly significant.
4. The point estimate of the odds ratio for the effect of SMK on cervical cancer status adjusted for the other variables in model I is given by $e^{1.4361} = 4.20$.

   The 95% interval estimate for the above odds ratio is given by

$$\exp\left[\hat{\beta}_{\text{SMK}} \pm 1.96\sqrt{\widehat{\text{Var}}\left(\hat{\beta}_{\text{SMK}}\right)}\right]$$

$$= \exp(1.4361 \pm 1.96 \times 0.3617) = \left(e^{0.8154}, e^{2.0568}\right)$$
$$= (2.26, 7.82).$$

5. Null hypothesis for the likelihood ratio test for the effect of SMK × NS: H_0: $\beta_{\text{SMK} \times \text{NS}} = 0$, where $\beta_{\text{SMK} \times \text{NS}}$ is the coefficient of SMK × NS in model II;

   Likelihood ratio statistic: $\text{LR} = -2\ln\ \hat{L}_{\text{I}} - (-2\ln\ \hat{L}_{\text{II}})$ where $\hat{L}_{\text{I}}$ and $\hat{L}_{\text{II}}$ are the maximized likelihood functions for models I and II, respectively. This statistic has approximately a chi-square distribution with one degree of freedom under the null hypothesis. Test computation: $\text{LR} = 174.97 - 171.46 = 3.51$. The P-value is less than 0.10 but greater than 0.05, which gives borderline significance because we would reject the null hypothesis at the 10% level but not at the 5% level. Thus, we conclude that the effect of the interaction of NS with SMK is of borderline significance.
6. Null hypothesis for the Wald test for the effect of SMK × NS is the same as that for the likelihood ratio test: H_0: $\beta_{\text{SMK} \times \text{NS}} = 0$ where $\beta_{\text{SMK} \times \text{NS}}$ is the coefficient of SMK × NS in model II;

Wald statistic: $Z = \frac{\hat{\beta}_{\mathrm{SMK} \times \mathrm{NS}}}{s_{\hat{\beta}_{\mathrm{SMK} \times \mathrm{NS}}}}$, which is approximately normal (0, 1) under H_0, or alternatively,
Z^2 is approximately chi square with one degree of freedom under H_0; test computation:
$Z = \frac{-1.1128}{0.5997} = -1.856$; alternatively, $Z^2 = 3.44$; the P-value for the Wald test is 0.0635, which gives borderline significance.
The LR statistic is 3.51, which is approximately equal to the square of the Wald statistic; therefore, both statistics give the same conclusion of borderline significance for the effect of the interaction term.

7. The formula for the estimated odds ratio is given by $\widehat{\mathrm{OR}}_{\mathrm{adj}} = \exp(\hat{\beta}_{\mathrm{SMK}} + \hat{\delta}_{\mathrm{SMK} \times \mathrm{NS}}\ \mathrm{NS}) = \exp(1.9381 - 1.1128\ \mathrm{NS})$, where the coefficients come from Model II and the confounding effects of NS and AS are controlled.

8. Using the adjusted odds ratio formula given in Question 7, the estimated odds ratio values for NS = 1 and NS = 0 are

   $\mathrm{NS} = 1$: $\exp[1.9381 - 1.1128(1)] = \exp(0.8253) = 2.28$;
   $\mathrm{NS} = 0$: $\exp[1.9381 - 1.1128(0)] = \exp(1.9381) = 6.95$

9. Formula for the 95% confidence interval for the adjusted odds ratio when NS = 1:

$$\exp\left[\hat{l} \pm 1.96\sqrt{\widehat{\mathrm{var}}(\hat{l})}\right],\ \text{where } \hat{l} = \hat{\beta}_{\mathrm{SMK}} + \hat{\delta}_{\mathrm{SMK} \times \mathrm{NS}}(1)$$
$$= \hat{\beta}_{\mathrm{SMK}} + \hat{\delta}_{\mathrm{SMK} \times \mathrm{NS}}$$

   and

$$\widehat{\mathrm{var}}(\hat{l}) = \widehat{\mathrm{var}}\left(\hat{\beta}_{\mathrm{SMK}}\right) + (1)^2\widehat{\mathrm{var}}\left(\hat{\delta}_{\mathrm{SMK} \times \mathrm{NS}}\right)$$
$$+ 2(1)\widehat{\mathrm{cov}}\left(\hat{\beta}_{\mathrm{SMK}}, \hat{\delta}_{\mathrm{SMK} \times \mathrm{NS}}\right),$$

   where $\widehat{\mathrm{var}}\left(\hat{\beta}_{\mathrm{SMK}}\right), \widehat{\mathrm{var}}\left(\hat{\delta}_{\mathrm{SMK} \times \mathrm{NS}}\right)$, and $\widehat{\mathrm{cov}}\left(\hat{\beta}_{\mathrm{SMK}}, \hat{\delta}_{\mathrm{SMK} \times \mathrm{NS}}\right)$ are obtained from the printout of the variance–covariance matrix.

10. $\hat{l} = \hat{\beta}_{\mathrm{SMK}} + \hat{\delta}_{\mathrm{SMK} \times \mathrm{NS}} = 1.9381 + (-1.1128) = 0.8253$
    $\widehat{\mathrm{var}}(\hat{l}) = 0.1859 + (1)^2(0.3596) + 2(1)(-0.1746)$
    $= 0.1859 + 0.3596 - 0.3492 = 0.1963.$

    The 95% confidence interval for the adjusted odds ratio is given by

$$\exp\left[\hat{l} \pm 1.96\sqrt{\widehat{\mathrm{Var}}(\hat{l})}\right] = \exp\left(0.8253 \pm 1.96\sqrt{0.1963}\right)$$
$$= \exp(0.8253 \pm 1.96 \times 0.4430)$$
$$= \left(e^{-0.0430}, e^{1.6936}\right) = (0.96, 5.44).$$

11. Model II is more appropriate than Model I if the test for the effect of interaction is viewed as significant. Otherwise, Model I is more appropriate than Model II. The decision here is debatable because the test result is of borderline significance.

Chapter 6

True–False Questions:

1. F: one stage is variable specification
2. T
3. T
4. F: no statistical test for confounding
5. F: validity is preferred to precision
6. F: for initial model, *V*s chosen a priori
7. T
8. T
9. F: model needs $E \times B$ also
10. F: list needs to include $A \times B$
11. The given model is hierarchically well formulated because for each variable in the model, every lower order component of that variable is contained in the model. For example, if we consider the variable SMK × NS × AS, then the lower order components are SMK, NS, AS, SMK × NS, SMK × AS, and NS × AS; all these lower order components are contained in the model.
12. A test for the term SMK × NS × AS is not dependent on the coding of SMK because the model is hierarchically well formulated and SMK × NS × AS is the highest-order term in the model.
13. A test for the terms SMK × NS is dependent on the coding because this variable is a lower order term in the model, even though the model is hierarchically well formulated.
14. In using a hierarchical backward elimination procedure, first test for significance of the highest-order term SMK × NS × AS, then test for significance of lower order interactions SMK × NS and SMK × AS, and finally assess confounding for *V* variables in the model. Based on the hierarchy principle, any two-factor product terms and *V* terms which are lower order components of higher order product terms found significant are not eligible for deletion from the model.
15. If SMK × NS × AS is significant, then SMK × NS and SMK × AS are interaction terms that must remain in

any further model considered. The V variables that must remain in further models are NS, AS, NS × AS, and, of course, the exposure variable SMK. Also the V^* variables must remain in all further models because these variables reflect the matching that has been done.

16. The model after interaction assessment is the same as the initial model. No potential confounders are eligible to be dropped from the model because NS, AS, and NS × AS are lower components of SMK × NS × AS and because the V^* variables are matching variables.

Chapter 7

1. The interaction terms are SMK × NS, SMK × AS, and SMK × NS × AS. The product term NS × AS is a V term, not an interaction term, because SMK is not one of its components.
2. Using a hierarchically backward elimination strategy, one would first test for significance of the highest-order interaction term, namely, SMK × NS × AS. Following this test, the next step is to evaluate the significance of two-factor product terms, although these terms might not be eligible for deletion if the test for SMK × NS × AS is significant. Finally, without doing statistical testing, the V variables need to be assessed for confounding and precision.
3. If SMK × NS is the only interaction found significant, then the model remaining after interaction assessment contains the V^* terms, SMK, NS, AS, NS × AS, and SMK × NS. The variable NS cannot be deleted from any further model considered because it is a lower order component of the significant interaction term SMK × NS. Also, the V^* terms cannot be deleted because these terms reflect the matching that has been done.
4. The odds ratio expression is given by $\exp(\beta + \delta_1\text{NS})$.
5. The odds ratio expression for the model that does not contain NS × AS has exactly the same form as the expression in Question 4. However, the coefficients β and δ_1 may be different from the Question 4 expression because the two models involved are different.
6. Drop NS × AS from the model and see if the estimated odds ratio changes from the gold standard model remaining after interaction assessment. If the odds ratio changes, then NS × AS cannot be dropped and is considered a confounder. If the odds ratio does not change, then NS × AS is not a confounder. However, it may still need to be controlled for precision reasons. To assess precision, one should compare confidence

intervals for the gold standard odds ratio and the odds ratio for the model that drops NS × AS. If the latter confidence interval is meaningfully narrower, then precision is gained by dropping NS × AS, so that this variable should, therefore, be dropped. Otherwise, one should control for NS × AS because no meaningful gain in precision is obtained by dropping this variable. Note that in assessing both confounding and precision, tables of odds ratios and confidence intervals obtained by specifying values of NS need to be compared because the odds ratio expression involves an effect modifier.

7. If NS × AS is dropped, the only *V* variable eligible to be dropped is AS. As in the answer to Question 6, confounding of AS is assessed by comparing odds ratio tables for the gold standard model and reduced model obtained by dropping AS. The same odds ratio expression as given in Question 5 applies here, where, again, the coefficients for the reduced model (without AS and NS × AS) may be different from the coefficient for the gold standard model. Similarly, precision is assessed similarly to that in Question 6 by comparing tables of confidence intervals for the gold standard model and the reduced model.
8. The odds ratio expression is given by exp(1.9381 – 1.1128NS). A table of odds ratios for different values of NS can be obtained from this expression and the results interpreted. Also, using the estimated variance–covariance matrix (not provided here), a table of confidence intervals (CIs) can be calculated and interpreted in conjunction with corresponding odds ratio estimates. Finally, the CIs can be used to carry out two-tailed tests of significance for the effect of SMK at different levels of NS.

Chapter 8

1. a. The screening approach described does not individually assess whether any of the control (C) variables are either potential confounders or potential effect modifiers.

   b. Screening appears necessary because the number of variables being considered (including possible interaction terms) for modeling is large enough to expect that either a model containing all main effects and interactions of interest won't run, or will yield unreliable estimated regression coefficients. In particular, if such a model does not run, collinearity assessment becomes difficult or even impossible, so the only way to get a "stable" model requires dropping some variables.

2. $$\begin{aligned}\text{logit } P(\mathbf{X}) = \alpha &+ \beta_1\text{F} + \beta_2\text{BASE} + \gamma_1\text{POST}\\ &+ \gamma_2\text{PF} + \gamma_3\text{OCC} + \gamma_4\text{AGE} + \gamma_5\text{GEN}\\ &+ \delta_1\text{F} \times \text{BASE} + \delta_2\text{F} \times \text{POST}\\ &+\delta_3\text{F} \times \text{PF} + \delta_4\text{F} \times \text{OCC} + \delta_5\text{F} \times \text{AGE}\\ &+\delta_6\text{F} \times \text{GEN} + \delta_7\text{BASE} \times \text{POST}\\ &+\delta_8\text{BASE} \times \text{PF}+ \delta_9\text{BASE} \times \text{OCC}\\ &+\delta_{10}\text{BASE} \times \text{AGE} + \delta_{11}\text{BASE} \times \text{GEN}\end{aligned}$$

3. **F × POST:** EV

   F × BASE: EE
   POST × BASE: EV
   PF × POST: Neither
   BASE × PF: EV

4. Choices *c* and *d* are reasonable.

   Choice *a* is not reasonable because the two nonsignificant chunk tests do not imply that the overall chunk test is nonsignificant; in fact, the overall chunk test was significant.
   Choice *b* is not reasonable because the corresponding chunk test was *more* significant than the chunk test for interaction terms involving BASE.

5. Using the model of question *2*,
   H_0: $\delta_2 = \delta_3 = \delta_4 = \delta_5 = \delta_6 = 0$.

   LR $= -2 \ln L_{\text{Reduced}} - (-2 \ln L_{\text{Full}}) \sim$ approx $\chi^2{}_5$ under H_0,
   where the Full Model is the model of question *2*, and the Reduced Model does not contain the product terms **F × POST, F × PF, F × OCC, F × AGE**, and **F×GEN**.

6. **PF, AGE,** and **GEN** are eligible to be dropped from the model. These variables are *V* variables that are not components of the three product terms found significant from interaction assessment.

7. No. The variables **POST** and **OCC** are lower order components of the product terms **F × POST** and **BASE × OCC** found significant, and are therefore to be retained in the model (from the hierarchy principle). Tests of significance for **POST** and **OCC** will depend on the coding of the variables in the model, so such tests are inappropriate, since they should be independent of coding.

8. a. $$\begin{aligned}\text{logit } P(\mathbf{X}) = \alpha &+ \beta_1\text{F} + \beta_2\text{BASE} + \gamma_1\text{POST} + \gamma_2\text{PF}\\ &+ \gamma_3\text{OCC} + \gamma_4\text{AGE} + \gamma_5\text{GEN}\\ &+ \delta_1\text{F} \times \text{BASE} + \delta_2\text{F} \times \text{POST}\\ &+ \delta_9\text{BASE} \times \text{OCC}\end{aligned}$$

   b. $\mathbf{X}^* = (3, 1, \text{POST}, \text{PF}, \text{OCC}, \text{AGE}, \text{GEN})$, whereas
   $\mathbf{X} = (2, 0, \text{POST}, \text{PF}, \text{OCC}, \text{AGE}, \text{GEN})$,
   so

   $$\text{OR} = \exp[\beta_1 + \beta_2 + 3\delta_1 + \delta_2\text{POST} + \delta_9\text{OCC}]$$

9. a. Seven subsets of possible confounders other than all confounders (in the gold standard model): {**AGE, GEN**}, {**AGE, PF**}, {**PF, GEN**}, {**AGE**}, {**GEN**}, {**PF**}, {no variables}.

b.

	OCC = 1	OCC = 0
POST = 1	$\widehat{\text{OR}}_{11}$	$\widehat{\text{OR}}_{10}$
POST = 0	$\widehat{\text{OR}}_{01}$	$\widehat{\text{OR}}_{00}$

$$\widehat{\text{OR}}_{ij} = \exp[\hat{\beta}_1 + \hat{\beta}_2 + 3\hat{\delta}_1 + \hat{\delta}_2\text{POST}_i + \hat{\delta}_9\text{OCC}_j]$$

c. The collection of ORs in a table controlling for a given subset of confounders would be compared to the corresponding table of ORs for the gold standard model. Those tables that collectively give essentially the "same" odds ratios as found in the gold standard table identify subsets of confounders that control for confounding. However, to decide whether one table of ORs is collectively the "same" as the gold standard, you typically will need to make a subjective decision, which makes this approach difficult to carry out.

d. You would construct two tables of confidence intervals, one for the gold standard model and the other for the reduced model obtained when **PF** and **GEN** are dropped from the model. Each table has the same form as shown above in part *9b*, except that confidence intervals would be put into each cell of the table instead of estimated ORs. You would then compare the two tables collectively to determine whether precision is gained (i.e., confidence interval width is smaller) when **PF** and **GEN** are dropped from the model.

10. a. Just because a model runs does not mean that there is no collinearity problem, but rather that there is no "perfect" collinearity problem (i.e, a "perfect" linear relationship among some predictor variables). If a collinearity problem exists, the estimated regression coefficients are likely to be highly unstable and therefore may give very misleading conclusions.

b. Focus first on the highest CNI of 97. Observe the VDPs corresponding to this CNI. Determine which variables are high (e.g., VDP > 0.5) and whether one of these variables (e.g., an interaction term) can be removed from the model. Rerun the reduced model to produce collinearity diagnostics (CNIs and VDPs) again, and proceed similarly until no collinearity problem is identified.

c. Yes. The high VDP value on **F × BASE** suggests that this product term is involved in a collinearity problem. Such a problem was not previously found or even considered when using SAS's LOGISTIC procedure to evaluate interaction. If it is decided that **F × BASE** is collinear with other terms, then it should be dropped from the model before any further modeling is carried out.

d. The "best" choice is iii.

11. a. Suggested strategy: For each subject in the dataset, compute DeltaBetas for the variables **F** and **BASE** in your initial model and in your final "best" model. Using plots of these DeltaBetas for each model, identify any subjects whose plot is "extreme" relative to the entire dataset. Do not use Cook's distance-type measures since such measures combine the influence of all variables in the model, whereas the study focus is on the effect of **F** and/or **BASE** variables. One problem with using DeltaBetas, however, is that such measures detect influence on a log OR scale rather than an OR $= \exp[\beta]$.

b. Any subject who is identified to be an influential observation may nevertheless be correctly measured on all predictor variables, so the researcher must still decide whether such a subject should be included in the study. A conservative approach is to drop from the data only those influential subjects whose measurements have errata that cannot be corrected.

12. There is no well-established method for reducing the number of tests performed when carrying out a modeling strategy to determine a "best" model. One approach is to drop from the model any collection of variables found to be not significant using a "chunk" test. Bonferroni-type corrections are questionable because the researcher does not know in advance how many tests will be performed.

Chapter 9

1. The data listing is in subject-specific (SS) format. Even though the data listing is not provided as part of the question, the fact that one of the predictors is a continuous variable indicates that it would not be convenient or useful to try to determine the distinct covariate patterns in the data.
2. There are 186 covariate patterns (i.e., unique profiles). The main reason for this is that the model contains the continuous variable AGE. If, instead, AGE was a binary variable, the model would only contain 2^4 or 16 covariate patterns.

3. No. The model is not fully parameterized, since it contains five parameters whereas the number of covariate patterns is 186.
4. No. The model does not perfectly predict the case/noncase status of each of the 289 subjects in the data.
5. a. The deviance value of 159.2017 is not calculated using the deviance formula
   $\text{Dev}(\hat{\boldsymbol{\beta}}) = -2\ln(\hat{L}_c/\hat{L}_{\max})$.

   In particular $-2\ln\hat{L}_c = 279.317$ and $-2\ln\hat{L}_{\max} = 0$, so $\text{Dev}(\hat{\boldsymbol{\beta}}) = 279.317$.

   b. The other logistic model, i.e., Model 2, is the model that is defined by the 186 covariate patterns that comprise the fully parameterized model derived from the four independent variables PREVHOSP, AGE, GENDER, and PAMU. This model will contain 185 predictor variables plus an intercept term.

   c. $159.2017 = -2\ln\hat{L}_{\text{Model 1}} - (-2\ln\hat{L}_{\text{Model 2}})$,
   where $-2\ln\hat{L}_{\text{Model 1}} = 279.3170$
   and $-2\ln\hat{L}_{\text{Model 2}} = 279.3170 - 159.2017 = 120.1153$.

   d. G = # of covariate patterns = 186 is large relative to $n = 289$.
6. a. The HL test has a *P*-value of 0.4553, which is highly nonsignificant. Therefore, the HL test indicates that the model does not have lack of fit.

   b. Models 1 and 2 as described in question 5b.

   c. Choose Model 1 since it provides adequate fit and is more parsimonious than Model 2. However, a LR test cannot be performed since the deviance for Model 1 is based on a large number of covariate patterns.

   d. Neither model perfectly fits the data, since neither model is a saturated model.
7. Consider the information shown in the output under the heading “Partition for the Hosmer and Lemeshow Test.”

   a. The 10 groups are formed by partitioning the 289 predicted risks into 10 deciles of risk, where, for example, the highest (i.e., 10th) decile contains approximately the highest 10% of the predicted risks.

   b. Because subjects with the same value for the predicted risk cannot be separated into different deciles.

   c. For group 5:
   Expected number of cases = sum of predicted risks for all 29 subjects in group 5.
   Expected number of noncases = 29 − expected number of cases

d. For group 5, term involving cases:
$(10-9.98)^2/9.98 = 0.00004$
term involving noncases:
$(20-20.02)^2/20.02 = 0.00002$

e. 20 terms in the sum, with 10 terms for cases (mrsa = 1) and 10 terms for noncases (mrsa = 0).

8. No. The model is not fully parameterized, since it contains eight parameters whereas the number of covariate patterns is 186.

9. No. The model does not perfectly predict the case/noncase status of each of the 289 subjects in the data.

10. a. The deviance value of 157.1050 is not calculated using the deviance formula

$$\text{Dev}(\hat{\boldsymbol{\beta}}) = -2\ln(\hat{L}_c/\hat{L}_{\max}).$$

In particular $-\,2\ln\hat{L}_c = 277.221$ and $-\,2\ln\hat{L}_{\max} = 0$, so $\text{Dev}(\hat{\boldsymbol{\beta}}) = 277.221$.

b. G = no. of covariate patterns = 186 is large relative to $n = 289$.

11. a. The HL test has a P-value of 0.9686, which is highly nonsignificant. Therefore, the HL test indicates that the model does not have lack of fit.

b. Although the Hosmer-Lemeshow test for the interaction model is more nonsignificant ($P = 0.9686$) than the Hosmer-Lemeshow test for the no-interaction model ($P = 0.4553$), both models adequately fit the data and these two P-values are not sufficient to conclude which model is preferable.

c. Perform a (LR) test of hypothesis that compares interaction and no-interaction models.
H_0: $\delta_1 = \delta_2 = \delta_2 = 0$ in the interaction model;

$$\begin{aligned}\text{LR} &= -2\ln\hat{L}_{\text{no interaction}} - (-2\ln\hat{L}_{\text{interaction}})\\ &= 279.317 - 277.221 = 2.096\\ &= \text{Dev}_{\text{Model 1}} - \text{Dev}_{\text{Model 2}} = 159.2017 - 157.1050\\ &= 2.0967,\end{aligned}$$

which has approximately a chi-square distribution with 3 d.f. under H_0. The P-value satisfies $0.50 < P < 0.60$, which indicates that H_0 should not be rejected. Conclusion: No-interaction model is preferred.

Chapter 10

1. a.

True (Observed) Outcome

	c_p **= 0.30**	$Y = 1$	$Y = 0$
Predicted	$Y = 1$	$n_{TP} = 98$	$n_{FP} = 68$
Outcome	$Y = 0$	$n_{FN} = 16$	$n_{TN} = 107$
		$n_1 = 114$	$n_0 = 175$

b. Sensitivity % = 100(98/114) = 86.0
Specificity % = 100(107/175) = 61.1
1 − specificity % = 100−61.1 = 100(68/175) = 39.9
False positive % = 100(68/166) = 41.0
False negative % = 100(16/123) = 13.0

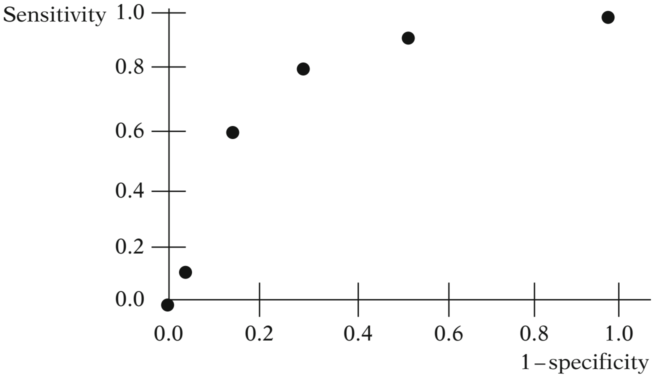
c. The denominator for calculating 1 − specificity is the number of true negatives (n_0 = 175) whereas the denominator for calculating the false positive percentage is 166, the total number of patients who were classified as postitive from the fitted model using the cut-point of 0.300.

d. Correct (%) = 100(98 + 107)/(114 + 175) = 70.9, which gives the percentage of all patients (i.e., cases and noncases combined) that were correctly classified as cases or noncases.

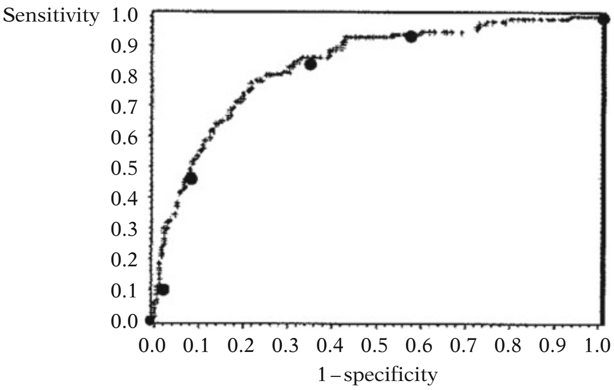
e. The sensitivity of 86.0% indicates that the model does well in predicting cases among the true cases. The specificity of 61.1 indicates that the model does not do very well in predicting noncases among the true noncases.

f. A drawback to assessing discrimination exclusively using the cut-point of 0.300 is that the sensitivity and specificity that results from a given cut-point may vary with the cut-point chosen.

2. Plots for the following cut-points: 0.000, 0.200, 0.400, 0.600, 0.800, and 1.000



3. a.



b. AUC = c = 0.840; good discrimination (grade B).

c. 19,950 = 114 × 175, where 114 is the number of true cases and 175 is the number of true noncases. Thus, 19,950 is the number of distinct case/noncase pairs in the dataset.

d. $$\text{AUC} = c = \frac{w + 0.5z}{n_{\text{p}}}$$
$$= \frac{19,950(.838) + 0.5(19,950)(.004)}{19,950} = 0.840$$

4. a. Area within entire rectangle = 114 × 715 = 19,950, which is the number of case/noncase pairs in the dataset used to compute the AUC.

b. The area under the superimposed ROC curve is 16,758, which is the numerator in the AUC formula, giving essentially the number of case/noncase pairs in which the predicted risk for the case is at least as large as the predicted risk for the corresponding noncase in the pair (where ties are weighed by 0.5).

5. a. Yes, the model fits the data because the HL statistic is highly nonsignificant. Also, the corresponding observed and expected cases are very close in each decile, as is the corresponding observed and expected noncases.

b. The distribution of observed cases indicates that the higher is the decile, the higher is the number of observed cases. The distribution of observed noncases indicates that the lower is the decile, the higher is the number of observed noncases. This indicates that the model provides good discrimination of the cases from the noncases.

c. The table provided for this part indicates that the model poorly discriminates cases from noncases and also provides poor GOF. Both the observed cases and the observed noncases appear to be uniformly distributed over the deciles. In contrast, good discrimination would be indicated by an increasing trend in the observed cases and a corresponding decreasing trend in the observed noncases as the deciles of predicted risk increase from low to high. The table indicates poor GOF since, when considered collectively, corresponding observed and expected cases differ substantially over most deciles; similarly corresponding observed and expected and noncases also differ substantially.

d. The table provided for this part indicates that the model discriminates poorly but provides good fit. Poor discrimination is indicated by the uniform distributions of both observed cases and noncases over the deciles. Good fit is indicated by the closeness of corresponding observed and expected cases and noncases over the deciles.

e. Yes, it is possible that a model might provide good discrimination but have poor fit. An example is given of data in which there is interaction, but a no-interaction model provides good discrimination despite not having good fit:

$V = 1$

	$E = 1$	$E = 0$
$D = 1$	13	12
$D = 0$	8	171

$V = 0$

	$E = 1$	$E = 0$
$D = 1$	38	70
$D = 0$	12	102

$$\widehat{\text{OR}}_{V=1} = 23.16 \qquad \widehat{\text{OR}}_{V=0} = 4.61$$

Model: logit $P(\mathbf{X}) = \beta_0 + \beta_1 E + \beta_2 V$

From the edited output shown below it can be seen that the AUC = 0.759, which indicates "fair" discrimination (Grade C), whereas the Hosmer GOF test indicates poor lack of fit ($P = 0.0416$):

```
              Analysis of Maximum Likelihood Estimates
                                                Wald
Parameter   DF   Estimate   Std Error      Chi-Sq    Pr > ChiSq
Intercept    1    -2.6484      0.2676     97.9353       <.0001
E            1     2.1026      0.3218     42.7024       <.0001
V            1     0.0872      0.3291      0.0702       0.7911

                     Odds Ratio Estimates
Effect      Pt Estimate       95%      Wald Confidence Limits
E                 8.188     4.358              15.383
V                 1.091     0.572               2.080
```

```
   Association of Predicted Probabilities and Observed Responses
Percent Concordant        65.2       Somers' D       0.518
Percent Discordant        13.4       Gamma           0.660
Percent Tied              21.5       Tau-a           0.144
Pairs                     25205      c               0.759

           Partition for the Hosmer and Lemeshow Test
                               Event                 Nonevent
Group   Total   Observed   Expected   Observed   Expected
1        179        8        11.83       171      167.17
2        114       12         8.17       102      105.83
3         25       13         9.17        12       15.83
4        108       38        41.83        70       66.17

           Hosmer and Lemeshow Goodness-of-Fit Test
Chi-Square                   DF                  Pr > ChiSq
  6.3576                      2                    0.0416
```

Chapter 11

True-False Questions:

1. T
2. F: information may be lost from matching: sample size may be reduced by not including eligible controls
3. T
4. T
5. T
6. McNemar's chi square: $(X - Y)^2/(X + Y) = (125 - 121)^2/(125 + 121) = 16/246 = 0.065$, which is highly nonsignificant. The MOR equals $X/Y = 125/121 = 1.033$. The conclusion from this data is that there is no meaningful or significant effect of exposure (Vietnam veteran status) on the outcome (genetic anomalies of offspring).
7. $\text{logit P}(\mathbf{X}) = \alpha + \beta\text{E} + \sum_{i=1}^{8501} \gamma_{1i}V_{1i}$,

   where the V_{1i} denote 8,501 dummy variables used to indicate the 8,502 matched pairs.
8. The Wald statistic is computed as $Z = 0.032/0.128 = 0.25$. The square of this Z is 0.0625, which is very close to the McNemar chi square of 0.065, and is highly nonsignificant.
9. The odds ratio from the printout is 1.033, which is identical to the odds ratio obtained using the formula X/Y.
10. The confidence interval given in the printout is computed using the formula

    $$\exp\left[\hat{\beta} \pm 1.96\sqrt{\widehat{\text{var}}\left(\hat{\beta}\right)}\right],$$

    where the estimated coefficient $\hat{\beta}$ is 0.032 and the square root of the estimated variance, i.e., $\sqrt{\widehat{\text{var}}(\hat{\beta})}$, is 0.128.

Chapter 12

True-False Questions:

1. F: The outcome categories are not ordered.
2. T
3. T
4. F: There will be four estimated coefficients for each independent variable.
5. F: The choice of reference category will affect the estimates and interpretation of the model parameters.
6. $$\begin{aligned}\text{Odds} &= \exp[\alpha_2 + \beta_{21}(40) + \beta_{22}(1) + \beta_{23}(0) + \beta_{24}(\text{HPT})]\\ &= \exp[\alpha_2 + 40\beta_{21} + \beta_{22} + (\text{HPT})\beta_{24}]\end{aligned}$$
7. $\text{OR} = \exp(\beta_{12})$
8. $\text{OR} = \exp[(50\text{–}20)\beta_{21}] = \exp(30\beta_{21})$
9. H_0: $\beta_{13} = \beta_{23} = \beta_{33} = \beta_{14} = \beta_{24} = \beta_{34} = 0$

   Test statistic: -2 log likelihood of the model without the smoking and hypertension terms (i.e., the reduced model), minus -2 log likelihood of the model containing the smoking and hypertension terms (i.e., the full model from Question 6).

   Under the null hypothesis the test statistic follows an approximate chi-square distribution with six degrees of freedom.
10. $$\ln\left[\frac{\text{P}(D=g|\mathbf{X})}{\text{P}(D=0|\mathbf{X})}\right] = \begin{aligned}[t][&\alpha_g + \beta_{g1}\text{AGE} + \beta_{g2}\text{GENDER}\\ &+\beta_{g3}\text{SMOKE} + \beta_{g4}\text{HPT}\\ &+\beta_{g5}(\text{AGE} \times \text{GENDER})\\ &+\beta_{g6}\text{GENDER} \times \text{SMOKE})],\end{aligned}$$

    where $g = 1, 2, 3$.

    Six additional parameters are added to the interaction model (β_{15}, β_{25}, β_{35}, β_{16}, β_{26}, β_{36}).

Chapter 13

True-False Questions:

1. T
2. T
3. F: each independent variable has one estimated coefficient
4. T
5. F: the odds ratio is invariant no matter where the cut-point is defined, but the *odds* is not invariant
6. T
7. F: the log odds ($D \geq 1$) is the log odds of the outcome being in category 1 or 2, whereas the log odds of $D \geq 2$ is the log odds of the outcome just being in category 2.
8. T: in contrast to linear regression, the actual values, beyond the order of the outcome variables, have no

effect on the parameter estimates or on which odds ratios are assumed invariant. Changing the values of the independent variables, however, may affect the estimates of the parameters.

9. odds = $\exp(\alpha_2 + 40\beta_1 + \beta_2 + \beta_3)$
10. OR = $\exp[(1-0)\,\beta_3 + (1-0)\,\beta_4] = \exp(\beta_3 + \beta_4)$
11. OR = $\exp(\beta_3 + \beta_4)$; the OR is the same as in Question 10 because the odds ratio is invariant to the cut-point used to dichotomize the outcome categories
12. OR = $\exp[-(\beta_3 + \beta_4)]$

Chapter 14

True-False Questions:

1. T
2. F: there is one common correlation parameter for all subjects.
3. T
4. F: a *function* of the mean response is modeled as linear (see next question)
5. T
6. T
7. F: only the estimated standard errors of the regression parameter estimates are affected. The regression parameter estimates are unaffected
8. F: consistency is an asymptotic property (i.e., holds as the number of clusters approaches infinity)
9. F: the empirical variance estimator is used to estimate the variance of the regression parameter estimates, not the variance of the response variable
10. T

Chapter 15

1. Model 1: $\text{logit } P(D=1\,|\,\mathbf{X}) = \beta_0 + \beta_1\text{RX} + \beta_2\text{SEQUENCE} + \beta_3\,\text{RX} \times \text{SEQ}$

   Model 2: $\text{logit } P(D=1\,|\,\mathbf{X}) = \beta_0 + \beta_1\text{RX} + \beta_2\text{SEQUENCE}$

   Model 3: $\text{logit } P(D=1\,|\,\mathbf{X}) = \beta_0 + \beta_1\text{RX}$
2. Estimated OR (where SEQUENCE = 0) = exp(0.4184) = 1.52

   95% CI: exp[0.4184 ± 1.96(0.5885)] = (0.48, 4.82)

   Estimated OR (where SEQUENCE = 1) = exp(0.4184 − 0.2136) = 1.23

   $$95\%\ \text{CI}: \exp\left[(0.4184 - 0.2136) \pm 1.96\sqrt{0.3463 + 0.6388 - 2(0.3463)}\,\right] = (0.43, 3.54)$$

Note: $\mathrm{var}(\hat{\beta}_1 + \hat{\beta}_3) = \mathrm{var}(\hat{\beta}_1) + \mathrm{var}(\hat{\beta}_3) + 2\ \mathrm{cov}(\hat{\beta}_1, \hat{\beta}_3)$. See Chap. 5.

3. The working covariance matrix pertains to the covariance between responses from the same cluster. The covariance matrix for parameter estimates pertains to the covariance between parameter estimates.
4. Wald test: H_0: $\beta_3 = 0$ for Model 1
   Test statistic: $z^2 = (0.2136/\ 0.7993)^2 = 0.069$;
   P-value = 0.79
   Conclusion: do not reject H_0.
5. Score test: H_0: $\beta_3 = 0$ for Model 1
   Test statistic = 0.07 test statistic distributed $\chi^2_{1\mathrm{df}}$
   P-value = 0.79
   Conclusion: do not reject H_0.
6. $\widehat{\mathrm{OR}}$ (from Model 2): $\exp(0.3104) = 1.36$;
   $\widehat{\mathrm{OR}}$ (from Model 3): $\exp(0.3008) = 1.35$.

   The odds ratios for RX are essentially the same whether SEQUENCE is or is not in the model, indicating that SEQUENCE is not a confounding variable by the data-based approach. From a theoretical perspective, SEQUENCE should not confound the association between RX and heartburn relief because the distribution of RX does not differ for SEQUENCE = 1 compared to SEQUENCE = 0. For a given patient's sequence, there is one observation where RX = 1, and one observation where RX = 0. Thus, SEQUENCE does not meet a criterion for confounding in that it is not associated with exposure status (i.e., RX).

Chapter 16

True-False Questions:

1. F: a marginal model does not include a subject-specific effect
2. T
3. T
4. T
5. T
6. $$\mathrm{logit}\ \mu_i = \beta_0 + \beta_1\mathrm{RX}_i + \beta_2\mathrm{SEQUENCE}_i + \beta_3\mathrm{RX}_i \times \mathrm{SEQ}_i + b_{0i}$$
7. $\widehat{\mathrm{OR}}(\text{where SEQUENCE}=0) = \exp(0.4707) = 1.60$
   $\widehat{\mathrm{OR}}(\text{where SEQUENCE}=1) = \exp(0.4707 - 0.2371) = 1.26$
8. $\widehat{\mathrm{OR}} = \exp(0.3553) = 1.43$
   $95\%\ \mathrm{CI}: \exp[0.3553 \pm 1.96(0.4565)] = (0.58, 3.49)$
9. The interpretation of the odds ratio, $\exp(\beta_1)$, using the model for this exercise is that it is the ratio of the odds for an individual (RX = 1 vs. RX = 0). The interpretation of the odds ratio for using a corresponding GEE

model (a marginal model) is that it is the ratio of the odds of a population average.

10. The variable SEQUENCE does not change values within a cluster since each subject has one specific sequence for taking the standard and active treatment. The matched strata are all concordant with respect to the variable SEQUENCE.

Bibliography

Anath, C.V., and Kleinbaum, D.G., Regression models for ordinal responses: A review of methods and applications, *Int. J. Epidemiol.* 26: 1323–1333, 1997.

Belsey, D.A., Kuh, E., and Welsch, R. E., *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York, NY, 1980.

Benjamini, Y., and Hochberg, Y., Controlling the false discovery rate: A practical and powerful approach to multiple testing, *J. R. Statist. Soc. B*, 57(1): 289–300, 1995.

Berkson, J. Limitations of the application of fourfold table analysis to hospital data, *Biometrics*, 2: 47–53, 1946.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W., *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, MA, 1975.

Breslow, N.E., and Clayton, D. G., Approximate inference in generalized linear mixed models, *J. Am. Stat. Assoc.* 88: 9–25, 1993.

Breslow, N.E., and Day, N.E., *Statistical Methods in Cancer Research, Vol. 1: The Analysis of Case-Control Studies*, IARC, Lyon, 1981.

*Brock, K.E., Berry, G., Mock, P. A., MacLennan, R., Truswell, A.S., and Brinton, L. A., Nutrients in diet and plasma and risk of in situ cervical cancer, *J. Natl. Cancer Inst.* 80(8):580–585, 1988.

Cannon, M.J., Warner, L., Taddei, J.A., and Kleinbaum, D.G., What can go wrong when you assume that correlated data are independent: An illustration from the evaluation of a childhood health intervention in Brazil, *Stat. Med.* 20: 1461–1467, 2001.

Carey, V., Zeger, S.L., and Diggle, P., Modelling multivariate binary data with alternating logistic regressions, *Biometrika* 80(3): 7–526, 1991.

Collett, D., *Modelling Binary Data*, Chapman and Hall, 1991.

*Sources for practice exercises or test questions presented at the end of several chapters.

*Donavan, J.W., MacLennan, R., and Adena, M., Vietnam service and the risk of congenital anomalies. A case-control study. *Med. J. Aust.* 149(7): 394–397, 1984.

Gavaghan, T.P., Gebski, V., and Baron, D.W., Immediate postoperative aspirin improves vein graft patency early and late after coronary artery bypass graft surgery. A placebo-controlled, randomized study, *Circulation* 83(5): 1526–1533, 1991.

Hill, H.A., Coates, R.J., Austin, H., Correa, P., Robboy, S.J., Chen, V., Click, L.A., Barrett, R.J., Boyce, J.G., Kotz, H.L., and Harlan, L.C., Racial differences in tumor grade among women with endometrial cancer, *Gynecol. Oncol*. 56: 154–163, 1995.

Hanley, J.A., McNeil, B.J., A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148(3): 839–843. http://radiology.rsnajnls.org/cgi/content/abstract/148/3/839, 1983

Hochberg, Y., A sharper Bonferroni procedure for multiple tests of significance, *Biometrika*, 75, 800–803, 1988.

Holm, S., A simple sequentially rejective multiple test procedure, *Scand. J. Stat.* 6: 65–70, 1979.

Kleinbaum, D.G., Kupper, L.L., and Chambless, L.E., Logistic regression analysis of epidemiologic data: Theory and practice, *Commun. Stat.* 11(5): 485–547, 1982.

Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H., *Epidemiologic Research: Principles and Quantitative Methods*, Wiley, New York, 1982.

Kleinbaum, D.G., Kupper, L.L., Nizam, A., and Muller, K.E., *Applied Regression Analysis and Other Multivariable Methods*, 4th ed., Duxbury Press/Cengage Learning, 2008.

Liang, K.Y., and Zeger, S.L., Longitudinal data analysis using generalized linear models, *Biometrika* 73: 13–22, 1986.

Light, R.J., and Pillemer, D.B., *Summing up. The Science of Reviewing Research*, Harvard University Press, Cambridge, MA, 1984.

Lipsitz, S.R., Laird, N.M., and Harrington, D.P., Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association, *Biometrika* 78(1): 153–160, 1991.

McCullagh, P., Regression models for ordinal data, *J. Roy. Stat. Soc.* B, 42(2): 109–142, 1980.

McCullagh, P., and Nelder, J.A., *Generalized Linear Models*, 2nd ed., Chapman & Hall, London, 1989.

*Sources for practice exercises or test questions presented at the end of several chapters.

*McLaws, N., Irwig, L.M., Mock, P., Berry, G., and Gold, J., Predictors of surgical wound infection in Australia: A national study, *Med. J. Aust.* 149: 591–595, 1988.

Pregibon, D., Logistic regression diagnostics, *Ann. Stat.* 9(4): 705–724, 1981.

Prentice, R.L., and Pyke, R., Logistic disease incidence models and case-control studies, *Biometrika* 66(3): 403–411, 1979.

Rezende, N.A., Blumberg, H.M., Metzger, B.S., Larsen, N.M., Ray, S.M., and McGowan, J.E., Risk factors for methicillin-resistance among patients with staphylococcus aureus bacteremia at the time of hospital admission, *Am. J. Med. Sci.*, 323 (3): 117–123, 2002.

Rothman, K.J., No adjustments are Needed for multiple comparisons, *Epidemiology*, 1(1): 43–46, 1990.

Rothman, K.J., Greenland, S., and Lash, T.L. *Modern Epidemiology*, 3rd ed., Lippincott Williams & Wilkins, Philadelphia, PA, 2008.

SAS Institute, *SAS/STAT User's Guide, Version 8.0*, SAS Institute, Inc., Cary, NC, 2000.

Sidack, Z., Rectangular confidence region for the means of multivariate normal distributions, *J. Am. Stat. Assoc.*, 62: 626–633, 1967.

Tigges, S., Pitts, S., Mukundan, S., Morrison, D., Olson, M., and Shahriara, A., External validation of the Ottawa Knee Rules in an Urban Trauma Center in the United States, *AJR*, 172: 1069–1071, 1999.

Wikipedia contributors. Receiver operating characteristic Wikipedia, The Free Encyclopedia. February 23, 2009, 20: 13 UTC. Available at: http://en.wikipedia.org/w/index.php?title=Receiver_operating_characteristic&oldid=272792033. Accessed March 16, 2009.

Wolfinger, R., and O'Connell, M., Generalized linear models: A pseudo-likelihood approach, *J. Stat. Comput. Simul.* 48: 233–243, 1993.

Zack, M., Singleton, J., and Satterwhite, C., Collinearity macro (SAS), Unpublished, Department of Epidemiology RSPH at Emory University, (contact dkleinb@sph.emory.edu), 2009.

*Sources for practice exercises or test questions presented at the end of several chapters.

Index

Springer
the language of science

springer.com

Survival Analysis

A Self-Learning Text

David G. Kleinbaum and Mitchel Klein

This greatly expanded second edition of Survival Analysis- A Self-learning Text provides a highly readable description of state-of-the-art methods of analysis of survival/event-history data.

Content: Introduction to Survival Analysis.- Kaplan-Meier Survival Curves and the Log-Rank Test.- The Cox Proportional Hazards Model and Its Characteristics.- Evaluating the Proportional Hazards Assumption.- The Stratified Cox Procedure.- Extension of the Cox Proportional Hazards Model for Time-Dependent Variables.- Parametric Survival Models.- Recurrent Events Survival Analysis.- Competing Risks Survival Analysis.

2005. 2nd ed. XVI, 590 p. 107 illus. Hardcover
Statistics for Biology and Health
ISBN: 978-0-387-23918-7

ActivEpi 2.0

David G. Kleinbaum

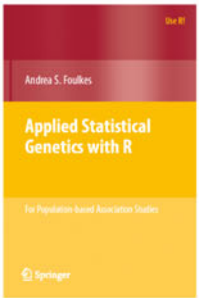
ActivEpi is a complete a multimedia presentation on CD-ROM of the material found in an introductory epidemiology course. Also, virtually all of the textual material on the ActivEpi CD-ROM is included in the ActivEpi Companion Textbook, which is also available for purchase from Springer.

Content: Introduction.- Epidemiologic research: An overview.- Epidemiologic study designs.- Measures of disease frequency.- Measure of effect.- Measures of potential impact.- Validity and general considerations.- Selection bias.- Information bias.- Confounding.- Confounding involving several risk factors.- Statistical inferences about effect measures.- Control of extraneous factors.- Stratified analysis.- Matching.

2008. Version 3.0 CD-ROM. Jewel case
ISBN: 978-0-387-77099-4

Applied Statistical Genetics with R For

Population-based Association Studies


Use R!
Andrea S. Foulkes
Applied Statistical Genetics with R
For Population-based Association Studies
Springer

Andrea S. Foulkes

This book presents fundamental concepts and principles in this emerging field at a level that is accessible to students and researchers with a first course in biostatistics. Extensive examples are provided using publicly available data and the open source, statistical computing environment, R.

Content: Genetic association studies.- Elementary statistical principles.- Genetic data concepts and tests.- Multiple comparison procedures.- Methods for unobservable phase.- Classification and regression trees.- Additional topics in high-dimensional data analysis.

2009. Approx. 270 p. Softcover
(Use R)
ISBN: 978-0-387-89553-6