

Statistics for Biology and Health

Series Editors

K. Dietz, M. Gail, K. Krickeberg, A. Tsiatis, J. Samet

Springer

New York

Berlin

Heidelberg

Hong Kong

London

Milan

Paris

Tokyo

Statistics for Biology and Health

Everitt/Rabe-Hesketh: Analyzing Medical Data Using S-PLUS

Ewens/Grant: Statistical Methods in Bioinformatics: An Introduction.

Hougaard: Analysis of Multivariate Survival Data.

Klein/Moeschberger: Survival Analysis: Techniques for Censored and Truncated Data.

Kleinbaum: Logistic Regression: A Self-Learning Text.

Kleinbaum/Klein: Logistic Regression: A Self-Learning Text, 2nd Ed.

Kleinbaum: Survival Analysis: A Self-Learning Text.

Lange: Mathematical and Statistical Methods for Genetic Analysis.

Manton/Singer/Suzman: Forecasting the Health of Elderly Populations.

Salsburg: The Use of Restricted Significance Tests in Clinical Trials.

Therneau/Grambsch: Modeling Survival Data: Extending the Cox Model.

Zhang/Singer: Recursive Partitioning in the Health Sciences

David G. Kleinbaum Mitchel Klein

Logistic Regression

A Self-Learning Text

Second Edition

With Contributions by
Erica Rihl Pryor



Springer

David G. Kleinbaum
Mitchel Klein
Department of Epidemiology
Emory University
Atlanta, GA 30333
USA

Series Editors

K. Dietz
Institut für Medizinische
Biometrie
Universität Tübingen
Westbahnhofstrasse 55
D-72070 Tübingen
Germany

M. Gail
National Cancer Institute
Rockville, MD 20892
USA

K. Krickeberg
Le Chatelet
F-63270 Manglieu
France

J. Samet
School of Public Health
Department of Epidemiology
Johns Hopkins University
615 Wolfe Street
Baltimore, MD 21205-2103
USA

A. Tsiatis
Department of Statistics
North Carolina State University
Raleigh, NC 27695
USA

Library of Congress Cataloging-in-Publication Data
Kleinbaum, David G.

Logistic regression: a self-learning text / David G. Kleinbaum,
Mitchel Klein.—2nd ed.
p. ; cm.—(Statistics for biology and health)
Includes bibliographical references and index.
ISBN 0-387-95397-3 (hc : alk. paper)
1. Medicine—Research—Statistical methods. 2. Regression
analysis. 3. Logistic distribution. I. Klein, Mitchel. II. Title.
III. Series
R853.S7 K54 2002
610'.7'27—dc21

2002019728

ISBN 0-387-95397-3

Printed on acid-free paper.

© 2002, 1994 Springer-Verlag New York, Inc.

All rights reserved. This work may not be translated or copied in whole or in part without the
written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New
York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly
analysis. Use in connection with any form of information storage and retrieval, electronic
adaptation, computer software, or by similar or dissimilar methodology now known or hereafter
developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even
if they are not identified as such, is not to be taken as an expression of opinion as to whether
or not they are subject to proprietary rights.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

SPIN 10858950

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science +Business Media GmbH

To John R. Boring III

Preface

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

As in the first edition, 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 and formulae 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 second edition has expanded the first edition by adding five new chapters and a new appendix. The five new chapters are

- Chapter 9. Polytomous Logistic Regression
- Chapter 10. Ordinal Logistic Regression
- Chapter 11. Logistic Regression for Correlated Data: GEE
- Chapter 12. GEE Examples
- Chapter 13. Other Approaches for Analysis of Correlated Data

Chapters 9 and 10 extend logistic regression to response variables that have more than two categories. Chapters 11–13 extend logistic regression to generalized estimating equations (GEE) and other methods for analyzing correlated response data.

The appendix is titled “Computer Programs for Logistic Regression” and provides descriptions and examples of computer programs for carrying out the variety of logistic regression procedures described in the main text. The software packages considered are SAS Version 8.0, SPSS Version 10.0, and STATA Version 7.0.

Also, Chapter 8 on the Analysis of Matched Data Using Logistic Regression has been expanded to include a discussion of three issues:

- Assessing interaction involving the matching variables
- Pooling exchangeable matched sets
- Analysis of matched follow-up data

Suggestions for Use

This text was originally intended for self-study, but in the eight years since the first edition was published it has also been effectively used as a text in a standard lecture-type classroom format. The text may be used to supplement material covered in a course or to review previously learned material in a self-instructional course or self-planned learning activity. 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 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 formula 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., Kupper, L., and Morgenstern, H., *Epidemiologic Research: Principles and Quantitative Methods*, John Wiley and Sons Publishers, New York, 1982, Chaps. 1–19.

Kleinbaum, D., Kupper, L., Muller, K., and Nizam, A., *Applied Regression Analysis and Other Multivariable Methods*, Third Edition, Duxbury Press, Pacific Grove, 1998, Chaps. 1–16.

A first course on the principles of epidemiologic research would be helpful as all modules in this series are written from the perspective of epidemiologic research. In particular, the learner should be familiar with the basic characteristics of epidemiologic study designs (follow-up, case control, and cross sectional) and should have some idea of the frequently encountered problem of controlling or adjusting for variables.

As for mathematics prerequisites, the reader 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, Georgia

David G. Kleinbaum
Mitchel Klein

Acknowledgments

David Kleinbaum and Mitch Klein wish to thank Erica Pryor at the School of Nursing, University of Alabama-Birmingham, for her many important contributions to this second edition. This includes fine-tuning the content of the five new chapters and appendix that have been added to the previous edition, taking primary responsibility for the pictures, formulae, symbols and summary information presented on the left side of the pages in each new chapter, performing a computer analysis of datasets described in the text, and carefully editing and correcting errata in the first eight chapters as well as the new appendix on computer software procedures.

All three of us (David, Mitch, and Erica) wish to thank John Boring, Chair of the Department of Epidemiology at the Rollins School of Public Health at Emory University for his leadership, inspiration, support, guidance, and friendship over the past several years. In appreciation, we are dedicating this edition to him.

Atlanta, GA

Birmingham, AL

David G. Kleinbaum

Mitchel Klein

Erica Rihl Pryor

Contents

Preface vii

Acknowledgments ix

Chapter 1

Introduction 2

Abbreviated Outline 2

Objectives 3

Presentation 4

Detailed Outline 29

Key Formulae 31

Practice Exercises 32

Test 34

Answers to Practice Exercises 37

Chapter 2

Important Special Cases of the Logistic Model 39

Introduction 40

Abbreviated Outline 40

Objectives 40

Presentation 42

Detailed Outline 65

Practice Exercises 67

Test 69

Answers to Practice Exercises 71

Chapter 3

Computing the Odds Ratio in Logistic Regression 73

Introduction 74

Abbreviated Outline 74

Objectives 75

Presentation 76

Detailed Outline 92

Practice Exercises 95

Test 97

Answers to Practice Exercises 99

Chapter 4

Maximum Likelihood Techniques: An Overview 101

Introduction 102

Abbreviated Outline 102

Objectives 103

Presentation	104
Detailed Outline	120
Practice Exercises	121
Test	122
Answers to Practice Exercises	124

Chapter 5

Statistical Inferences Using Maximum Likelihood Techniques 125

Introduction	126
Abbreviated Outline	126
Objectives	127
Presentation	128
Detailed Outline	150
Practice Exercises	152
Test	156
Answers to Practice Exercises	158

Chapter 6

Modeling Strategy Guidelines 161

Introduction	162
Abbreviated Outline	162
Objectives	163
Presentation	164
Detailed Outline	183
Practice Exercises	184
Test	186
Answers to Practice Exercises	188

Chapter 7

Modeling Strategy for Assessing Interaction and Confounding 191

Introduction	192
Abbreviated Outline	192
Objectives	193
Presentation	194
Detailed Outline	221
Practice Exercises	222
Test	224
Answers to Practice Exercises	225

Chapter 8**Analysis of Matched Data Using Logistic Regression 227**

Introduction 228
Abbreviated Outline 228
Objectives 229
Presentation 230
Detailed Outline 243
Practice Exercises 245
Test 247
Answers to Practice Exercises 249

Chapter 9**Polytomous Logistic Regression 267**

Introduction 268
Abbreviated Outline 268
Objectives 269
Presentation 270
Detailed Outline 293
Practice Exercises 295
Test 297
Answers to Practice Exercises 298

Chapter 10**Ordinal Logistic Regression 301**

Introduction 302
Abbreviated Outline 302
Objectives 303
Presentation 304
Detailed Outline 320
Practice Exercises 322
Test 324
Answers to Practice Exercises 325

Chapter 11**Logistic Regression for Correlated Data: GEE 327**

Introduction 328
Abbreviated Outline 328
Objectives 329
Presentation 330
Detailed Outline 367
Practice Exercises 373
Test 374
Answers to Practice Exercises 375

Chapter 12

GEE Examples 377

Introduction 378
Abbreviated Outline 378
Objectives 379
Presentation 380
Detailed Outline 396
Practice Exercises 397
Test 400
Answers to Practice Exercises 402

Chapter 13

Other Approaches for Analysis of Correlated Data 405

Introduction 406
Abbreviated Outline 406
Objectives 407
Presentation 408
Detailed Outline 427
Practice Exercises 429
Test 433
Answers to Practice Exercises 435

Appendix: Computer Programs for Logistic Regression 437

Datasets 438
SAS 441
SPSS 464
Stata 474

Test Answers 487

Bibliography 503

Index 507

1

Introduction to Logistic Regression

Contents	
Introduction	2
Abbreviated Outline	2
Objectives	3
Presentation	4
Detailed Outline	29
Key Formulae	31
Practice Exercises	32
Test	34
Answers to Practice Exercises	37

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.

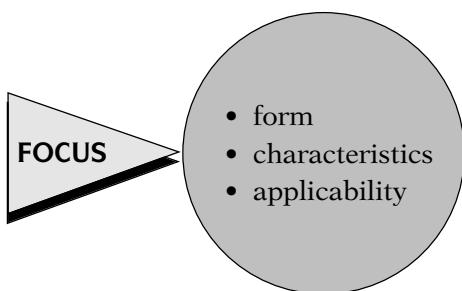
- I. The multivariable problem (pages 4–5)
- II. Why is logistic regression popular? (pages 5–7)
- III. The logistic model (pages 7–8)
- IV. Applying the logistic model formula (pages 9–11)
- V. Study design issues (pages 11–15)
- VI. Risk ratios versus odds ratios (pages 15–16)
- VII. Logit transformation (pages 16–22)
- VIII. Derivation of OR formula (pages 22–25)
- IX. Example of OR computation (pages 25–26)
- X. Special case for (0, 1) variables (pages 27–28)

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 follow-up, 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 \mathbf{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



EXAMPLE

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

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

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

“control for”

$$C_1 = \text{AGE}$$

$$C_2 = \text{RACE}$$

$$C_3 = \text{SEX}$$

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)?

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 .

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



Independent variables:

$$X_1, X_2, \dots, X_k$$

X 's may be E 's, C 's, or combinations

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.

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

EXAMPLE

$X_1 = E$	$X_4 = E \times C_1$
$X_2 = C_1$	$X_5 = C_1 \times C_2$
$X_3 = C_2$	$X_6 = E^2$

The Multivariable Problem

$$X_1, X_2, \dots, X_k \rightarrow D$$

The analysis:
mathematical model

Logistic model:
dichotomous D

Logistic is most popular

For example, we may have:

- 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

Whenever we wish to relate a set of X 's 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.

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

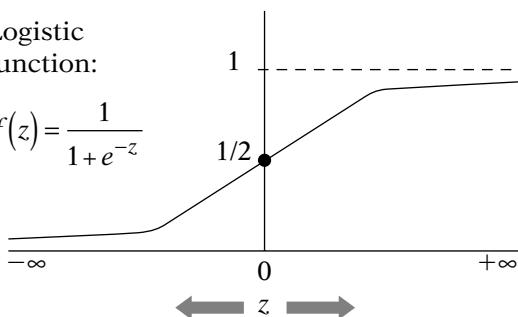
Other modeling approaches are possible also, but logistic regression is by far the most **popular** modeling procedure used to analyze epidemiologic data when the illness measure is dichotomous. We will show why this is true.

II. Why Is Logistic Regression Popular?

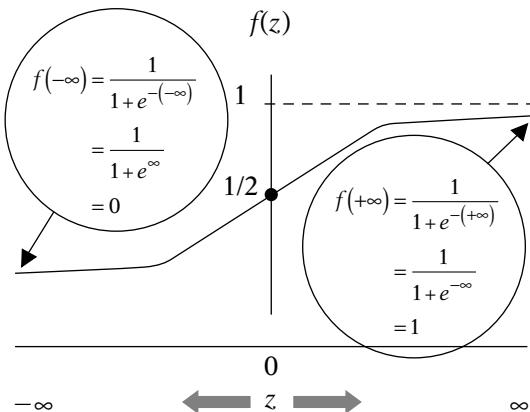
Logistic function:

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

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 $+\infty$.



6 1. Introduction to Logistic Regression



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

$0 \leq \text{probability} \leq 1$
(individual risk)

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

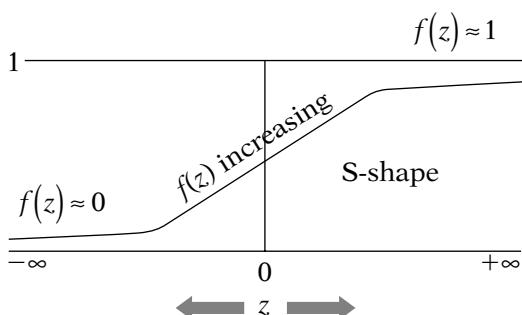
On the right side, when z is $+\infty$, then $f(z)$ equals 1.

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

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.

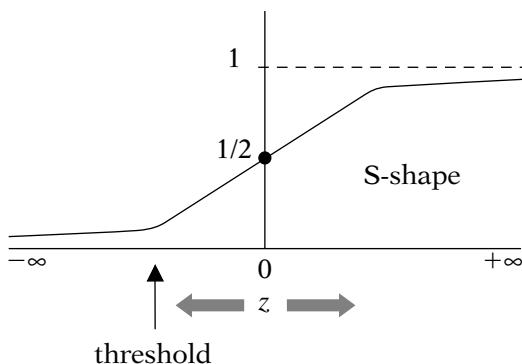
The **logistic model**, therefore, is set up to ensure that whatever estimate of risk we get, it will always be some number between 0 and 1. Thus, for the logistic model, we can never get a risk estimate either above 1 or below 0. This is not always true for other possible models, which is why the logistic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

- 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

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

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

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 X 's are independent variables of interest and α and the β_i are constant terms representing unknown parameters.

In essence, then, z is an index that combines the X 's.

$$\begin{aligned} z &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k \\ f(z) &= \frac{1}{1 + e^{-z}} \\ &= \frac{1}{1 + e^{-(\alpha + \sum \beta_i X_i)}} \end{aligned}$$

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, \dots, X_k measured at T_0

Time: $T_0 \rightarrow T_1$

$X_1, X_2, \dots, X_k \rightarrow D_{(0, 1)}$

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

DEFINITION

Logistic model:

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

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

↑ ↑

unknown parameters

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 .

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

The model is defined as **logistic** if the expression for the probability of developing the disease, given the X 's, 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 X 's 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.

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

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 $\beta_i X_i$.

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

Model formula:

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

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

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

DEFINITION

fit: use data to estimate

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

NOTATION

hat= $\hat{}$

parameter \Leftrightarrow estimator

$$\alpha \beta_1 \beta_2 \quad \hat{\alpha} \hat{\beta}_1 \hat{\beta}_2$$

Method of estimation:

maximum likelihood (ML)—
see Chapters 4 and 5

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.

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 .

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 $\hat{\alpha}$ “hat,” $\hat{\beta}_1$ “hat,” $\hat{\beta}_2$ “hat,” and $\hat{\beta}_3$ “hat.”

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

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

EXAMPLE

$$\hat{\alpha} = -3.911$$

$$\hat{\beta}_1 = 0.652$$

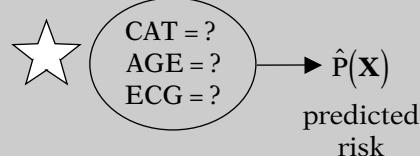
$$\hat{\beta}_2 = 0.029$$

$$\hat{\beta}_3 = 0.342$$

EXAMPLE (continued) $\hat{P}(\mathbf{X})$

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

$\hat{P}(\mathbf{X}) = ?$



CAT = 1
AGE = 40
ECG = 0

$$\begin{aligned}\hat{P}(\mathbf{X}) &= \frac{1}{1 + e^{-[-3.911 + 0.652(1) + 0.029(40) + 0.342(0)]}} \\ &= \frac{1}{1 + e^{-(2.101)}} \\ &= \frac{1}{1 + 8.173} \\ &= 0.1090\end{aligned}$$

CAT = 1
AGE = 40
ECG = 0 CAT = 0
AGE = 40
ECG = 0

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

11% risk / 6% risk

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} on the left-hand side of the formula because our estimated model will give us an estimated probability, not the exact probability.

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

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} for this individual. This estimate is often called a “predicted risk,” or simply “risk.”

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

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%.

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.

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

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

- RR (direct method)

Conditions for RR (direct method)

- ✓ follow-up study
- ✓ specify all X 's

- RR (indirect method)

- ✓ OR
- ✓ assumptions

- OR: direct estimate from

- ✓ follow-up
- ✓ case-control
- ✓ cross-sectional

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 $\widehat{\text{RR}}$, 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.

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.

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.

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.

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

$$X_1, X_2, \dots, X_k \rightarrow D_{(0, 1)}$$

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

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

robust conditions
case-control studies

robust conditions
cross-sectional studies

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

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:



Treat case control like 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.

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

Risk: only in follow-up

OR: case-control or cross-sectional

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

Case-control and cross-sectional studies:

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

$$\begin{array}{l} \hat{P}(E=1 | D=1) \xrightarrow{\quad} P(E|D) \text{ (general form)} \\ \hat{P}(E=1 | D=0) \xrightarrow{\quad} \end{array}$$

Risk: $P(D|E)$ 

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

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.

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 case-control or cross-sectional study.

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

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.

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

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

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 $P(D|E)$.

Case-control or cross-sectional studies:

~~P(D|E)~~
✓ $P(E|D) \Rightarrow$ risk

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

Case control:

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

Follow-up:

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

Case-control and cross-sectional:

✓ $\hat{\beta}_i$, \widehat{OR}

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

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 β 's in the logistic model. To get a predicted risk $\hat{P}(\mathbf{X})$ from fitting this model, we must obtain valid estimates of α and the β 's, these estimates being denoted by "hats" over the parameters in the mathematical formula for the model.

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 $\hat{P}(\mathbf{X})$ because $\hat{\alpha}$ is required for the computation.

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

Now, although α cannot be estimated from a case-control or cross-sectional study, the β 's can be estimated from such studies. As we shall see shortly, the β 's 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.

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

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

~~alpha~~

EXAMPLE (repeated)

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

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

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

Model :

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

EXAMPLE (continued)

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

AGE unspecified but fixed

ECG unspecified but fixed

Nevertheless, it is more common to obtain an estimate of a risk ratio or odds ratio without explicitly specifying the control variables. In our example, for instance, it is typical 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{OR} directly \widehat{RR} indirectly
provided $\widehat{OR} \approx \widehat{RR}$ $\widehat{OR} \approx \widehat{RR}$ if rare disease

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 odds ratio **directly**, but we cannot estimate the risk ratio. We can, however, stretch our interpretation to obtain a risk ratio **indirectly** provided we are willing to make certain assumptions. The key **assumption** here is that the odds ratio provides a good approximation to the risk ratio.

From previous exposure to epidemiologic principles, you may recall that one way to justify an odds ratio approximation for a risk ratio is to assume that the disease is rare. Thus, if we invoke the **rare disease assumption**, we can assume that the odds ratio estimate from a logistic regression model approximates a risk ratio.

Rare disease		OR	RR
yes	✓	✓	
no	✓	?	

If we cannot invoke the rare disease assumption, we cannot readily claim that the odds ratio estimate obtained from logistic modeling approximates a risk ratio. The investigator, in this case, may have to review the specific characteristics of the study before making a decision. It may be necessary to conclude that the odds ratio is a satisfactory measure of association in its own right for the current study.

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.

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

where

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

- (1) $P(\mathbf{X})$
- (2) $1-P(\mathbf{X})$
- (3) $\frac{P(\mathbf{X})}{1-P(\mathbf{X})}$
- (4) $\ln_e \left[\frac{P(\mathbf{X})}{1-P(\mathbf{X})} \right]$

EXAMPLE

- (1) $P(\mathbf{X}) = 0.110$
- (2) $1-P(\mathbf{X}) = 0.890$
- (3) $\frac{P(\mathbf{X})}{1-P(\mathbf{X})} = \frac{0.110}{0.890} = 0.123$
- (4) $\ln_e \left[\frac{P(\mathbf{X})}{1-P(\mathbf{X})} \right] = \ln(0.123) = -2.096$
i.e., $\text{logit}(0.110) = -2.096$

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

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

- (1) computing $P(\mathbf{X})$ and
- (2) 1 minus $P(\mathbf{X})$ separately, then
- (3) dividing one by the other, and finally
- (4) taking the natural log of the ratio.

For example, if $P(\mathbf{X})$ is 0.110, then

1 minus $P(\mathbf{X})$ 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 .

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 $P(\mathbf{X})$, which is 1 over 1 plus e to minus the quantity α plus the sum of the $\beta_i X_i$.

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

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

The diagram shows two mathematical expressions side-by-side. The first expression is $\text{logit } P(\mathbf{X}) = \ln_e \left[\frac{P(\mathbf{X})}{1-P(\mathbf{X})} \right]$. The second expression is $P(\mathbf{X}) = \frac{1}{1+e^{-(\alpha+\sum \beta_i X_i)}}$. There are three arrows originating from the term $P(\mathbf{X})$ in the first expression and pointing to the term $P(\mathbf{X})$ in the second expression. One arrow points to the numerator $P(\mathbf{X})$, another to the denominator $1-P(\mathbf{X})$, and the third to the exponent $-(\alpha+\sum \beta_i X_i)$.

18 1. Introduction to Logistic Regression

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

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

$$\frac{P(\mathbf{X})}{1 - P(\mathbf{X})} = \frac{\overbrace{1 + e^{-(\alpha + \sum \beta_i X_i)}}^{\text{e to minus the quantity } \alpha \text{ plus the sum of } \beta_i X_i}}{\overbrace{e^{-(\alpha + \sum \beta_i X_i)}}^{\text{one over 1 plus } e to minus } \alpha \text{ plus the sum of } \beta_i X_i}$$

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

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

$$= \underbrace{(\alpha + \sum \beta_i X_i)}_{\text{linear sum}}$$

Logit form:

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

where

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

logit $P(\mathbf{X})$  OR

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

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

Also, using some algebra, we can write $1 - P(\mathbf{X})$ as:

e to minus the quantity α plus the sum of $\beta_i X_i$ divided by one over 1 plus e to minus α plus the sum of the $\beta_i X_i$.

If we divide $P(\mathbf{X})$ by $1 - P(\mathbf{X})$, then the denominators cancel out,

and we obtain e to the quantity α plus the sum of the $\beta_i X_i$.

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

the linear sum α plus the sum of $\beta_i X_i$.

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

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

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 $P(\mathbf{X})$ divided by $1 - P(\mathbf{X})$, whose log value gives the **logit**, describes the **odds** for developing the disease for a person with independent variables specified by \mathbf{X} .

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

$$P = 0.25$$

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

$\frac{1}{3} \leftarrow$ event occurs

$\frac{2}{3} \leftarrow$ event does not occur

3 to 1 event will not happen

$$\text{odds} : \left[\frac{P(\mathbf{X})}{1-P(\mathbf{X})} \right] \text{ vs. } \frac{P}{1-P}$$

describes risk in logistic model for individual \mathbf{X}

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

= log odds for individual \mathbf{X}

$$= \alpha + \sum \beta_i X_i$$

EXAMPLE

$$\text{all } X_i = 0: \text{logit } P(\mathbf{X}) = ?$$

↓ 0

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

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

INTERPRETATION

- (1) $\alpha = \text{log odds for individual with all } X_i = 0$

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 $P(\mathbf{X})$ divided by $1-P(\mathbf{X})$ has essentially the same interpretation as P over $1-P$, which ignores \mathbf{X} .

The main difference between the two formulae is that the expression with the \mathbf{X} is more specific. That is, the formula with \mathbf{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 \mathbf{X} , and that we are interested in the odds associated with a particular specification of \mathbf{X} .

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 X 's.

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

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

If we plug in 0 for all the X 's in the formula, we find that the logit of $P(\mathbf{X})$ 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 .

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

EXAMPLE (continued)(2) $\alpha = \log$ of background odds**LIMITATION OF (1)**All $X_i = 0$ for any individual?

↓

 $AGE \neq 0$ $WEIGHT \neq 0$ (2) $\alpha = \log$ of background odds**DEFINITION OF (2)**background odds: ignores all X 's

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

$$\begin{matrix} \alpha & \checkmark \\ \beta_i? \end{matrix}$$

 $X_1, X_2, \dots, X_i, \dots, X_k$
 fixed varies fixed
EXAMPLECAT changes from 0 to 1;
AGE = 40, ECG = 0

fixed

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

A second interpretation is that α gives the **log of the background, or baseline, odds**.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.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.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.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.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.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.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.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
 $\text{logit } P(\mathbf{X}) = \alpha + \beta_1 + \beta_2 40 + \beta_3 0$
 $= (\alpha + \beta_1 + 40\beta_2)$

(2) CAT = 0, AGE = 40, ECG = 0
 $\text{logit } P(\mathbf{X}) = \alpha + \beta_1 0 + \beta_2 40 + \beta_3 0$
 $= (\alpha + 40\beta_2)$

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

NOTATION

Δ = change

$\beta_1 = \Delta \text{ logit}$ when $\Delta \text{ CAT} = 1$
 $= \Delta \text{ log odds}$ AGE and ECG fixed

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

$i = L$:

$$\beta_L = \Delta \ln(\text{odds})$$

when $= \Delta X_L = 1$, other X 's fixed

SUMMARY

α = background log odds	β_i = change in log odds
--------------------------------	--------------------------------

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

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

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.

Thus, letting the symbol Δ 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.

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

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.

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*.

logit  OR

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

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

EXAMPLE

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

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.

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

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) **X**₁ = (X₁₁, X₁₂, ..., X_{1k})
- (0) **X**₀ = (X₀₁, X₀₂, ..., X_{0k})

EXAMPLE

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

- (1) **X**₁ = (CAT = 1, AGE = 40, ECG = 0)
- (0) **X**₀ = (CAT = 0, AGE = 40, ECG = 0)

Let **X**₁ denote the collection of *X*'s that specify group 1 and let **X**₀ denote the collection of *X*'s that specify group 0.

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

X is the collection of variables CAT, AGE, and ECG,
X₁ corresponds to CAT=1, AGE=40, and ECG=0,
 whereas
X₀ 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 **X**₁ and **X**₀ in an **odds ratio**, we can write **OR**_{X₁, X₀} equals the **odds for X₁** divided by the **odds for X₀**.

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

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

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

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

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

$$\text{ROR} = \frac{\frac{P(\mathbf{X}_1)}{1 - P(\mathbf{X}_1)}}{\frac{P(\mathbf{X}_0)}{1 - P(\mathbf{X}_0)}}$$

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

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

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

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

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

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

Given a logistic model of the general form $P(\mathbf{X})$,

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

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

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.

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 $P(\mathbf{X})$ 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.

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.

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 .

24 1. Introduction to Logistic Regression

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

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

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

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

In computing this difference, the α 's cancel out and the β_i 's can be factored for the i th variable.

Thus, the expression for **ROR** simplifies to the quantity e to the sum β_i times the difference between X_{1i} and 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 \mathbf{X}_1 and \mathbf{X}_0 . Note that the formula involves the β_i 's but not α .

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 .

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.

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 $(X_{1i} - X_{0i})$,

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

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

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

-

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

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

$$e^{\sum_{i=1}^k z_i} = e^{z_1} \times e^{z_2} \times \dots \times e^{z_k}$$

NOTATION

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



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

-

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

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

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

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

- 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

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 \Rightarrow 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**.

A model different from the logistic model, depending on its form, might imply a different (for example, 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

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

EXAMPLE

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

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

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.

For example, suppose the \mathbf{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.

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

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

EXAMPLE (continued)

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

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

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

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

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

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

$$(1) \begin{pmatrix} \text{CAT} = 1, \text{AGE} = 40, \text{ECG} = 0 \\ \text{CAT} = 0, \text{AGE} = 40, \text{ECG} = 0 \end{pmatrix}$$

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

= an “adjusted” OR

AGE and ECG:

- fixed
- same
- control variables

e^{β_1} : population ROR

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

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

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.

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).

The last two terms reduce to 0,

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.

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.

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**.

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

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

EXAMPLE

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

adjusted

$X_i(0, 1)$: adj. ROR = e^{β_i}

controlling for other X 's

EXAMPLE

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

adjusted

ECG (0, 1): adj. ROR = e^{β_3}

controlling for CAT and AGE

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.

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

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.

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 :

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

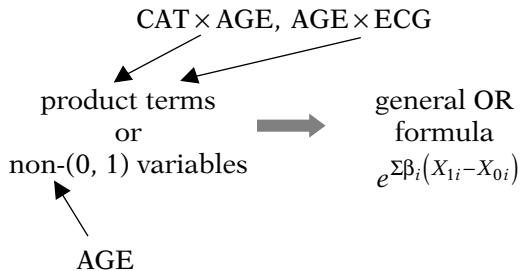
Thus, we can obtain an adjusted odds ratio for each (0, 1) variable in the logistic model by exponentiating the coefficient corresponding to that variable. This formula is much simpler than the general formula for ROR described earlier.

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 \times AGE or AGE \times ECG.



When the model contains product terms, like $CAT \times AGE$, 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 is 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: $P(D = 1 | X_1, \dots, X_k) = P(\mathbf{X}) = 1/[1 + \exp[-(\alpha + \sum \beta_i X_i)]]$.

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 $\hat{\alpha} = -3.911$, $\hat{\beta}_1(\text{CAT}) = 0.65$, $\hat{\beta}_2(\text{AGE}) = 0.029$, and $\hat{\beta}_3(\text{ECG}) = 0.342$.
- C. Predicted risk computations:
 $\hat{P}(\mathbf{X})$ for CAT=1, AGE=40, ECG=0: 0.1090,
 $\hat{P}(\mathbf{X})$ for CAT=0, AGE=40, 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 versus 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? May need to review characteristics of study to decide if the computed OR approximates an RR.

VII. Logit transformation (pages 16–22)

- A. Definition of the logit transformation:

$$\text{logit } P(\mathbf{X}) = \ln_e[P(\mathbf{X})/(1-P(\mathbf{X}))].$$
- B. The formula for the logit function in terms of the parameters of the logistic model: $\text{logit } P(\mathbf{X}) = \alpha + \sum \beta_i X_i$.
- 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 \mathbf{X} .
 - ii. The logit function describes the “log odds” for a person or group specified by \mathbf{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 X 's 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 X 's for groups 1 and 0.
- B. Example involving CAT, AGE, and ECG variables:
 $\mathbf{X}_1=(\text{CAT}=1, \text{AGE}=40, \text{ECG}=0), \mathbf{X}_0=(\text{CAT}=0, \text{AGE}=40, \text{ECG}=0).$

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

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

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

$$\text{ROR} = \exp[\sum \beta_i (X_{1i} - X_{0i})] = \prod \{\exp[\beta_i (X_{1i} - X_{0i})]\}$$

- E. Interpretation from the product (\prod) 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 X_1 and X_0 specified in VIII B above:

$\text{ROR} = \exp(\beta_1)$, where β_1 is the coefficient of CAT.

- B. Interpretation of $\exp(\beta_1)$: 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 X 's in model is given by the formula $\text{ROR} = \exp(\beta_i)$, 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 X 's, 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: $\text{logit } P(\mathbf{X}) = \alpha + \sum \beta_i X_i$

RISK ODDS RATIO (general formula):

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

RISK ODDS RATIO [(0, 1) variables]: $\text{ROR} = \exp(\beta_i)$ for the effect of the variable X_i *adjusted* for the other X 's

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		

- 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:

- 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 don't 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 don't 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 aren't 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.
- $\exp(\beta_S)$, where β_S is the coefficient of SOC in model 1.
 - $\exp[\sum \beta_i(X_{1i} - X_{0i})]$.
 - $\Pi[\exp[\beta_i(X_{1i} - X_{0i})]]$.

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 provided 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 $SMK \times RACE$ and $SMK \times AGE$, 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.)
- The model form can be written as $P(\mathbf{X}) = 1/[1 + \exp[-(\alpha + \sum \beta_i X_i)]]$, where “ $\exp[\cdot]$ ” denotes the quantity e raised to the power of the expression inside the brackets.
 - logit $P(\mathbf{X}) = \alpha + \sum \beta_i X_i$ is an alternative way to state the model.
 - $ROR = \exp[\sum \beta_i (X_{1i} - X_{0i})]$ is a general expression for the odds ratio that compares two groups of \mathbf{X} variables.
 - $ROR = \prod \{\exp[\beta_i (X_{1i} - X_{0i})]\}$ is a general expression for the odds ratio that compares two groups of \mathbf{X} variables.
 - 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:* $\hat{P}(\mathbf{X}) = 1/(1 + \exp\{-[-1.18 - 0.52(\text{SOC}) + 0.04(\text{SBP}) - 0.56(\text{SMK}) - 0.033(\text{SOC} \times \text{SBP}) + 0.175(\text{SOC} \times \text{SMK})]\})$.

Model 2: $\hat{P}(\mathbf{X}) = 1/(1 + \exp\{-[-1.19 - 0.50(\text{SOC}) + 0.01(\text{SBP}) + 0.42(\text{SMK})]\})$.

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

Model 2: $\text{logit } \hat{P}(\mathbf{X}) = -1.19 - 0.50(\text{SOC}) + 0.01(\text{SBP}) - 0.42(\text{SMK})$.

3. For SOC=1, SBP=150, and SMK=1,
 $\mathbf{X} = (\text{SOC}, \text{SBP}, \text{SMK}, \text{SOC} \times \text{SBP}, \text{SOC} \times \text{SMK}) = (1, 150, 1, 150, 1)$ and

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

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

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

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

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

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=150, 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,

$$\text{RR}(1 \text{ vs. } 2) = \frac{P(\text{SOC}=0, \text{SMK}=1, \text{SBP}=150)}{P(\text{SOC}=1, \text{SMK}=1, \text{SBP}=150)}$$

$$= 0.352/0.473 = 1/1.34 = 0.744$$

This estimated risk ratio is less than 1 because the risk for high social class persons ($\text{SOC}=1$) is less than the risk for low social class persons ($\text{SOC}=0$) 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. $\widehat{\text{OR}}(\text{SOC}=1 \text{ vs. } \text{SOC}=0 \text{ controlling for SMK and SBP})$

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

$$= \exp(-0.50) \\ = 0.6065 = 1/1.65.$$

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	
Introduction	40
Abbreviated Outline	40
Objectives	40
Presentation	42
Detailed Outline	65
Practice Exercises	67
Test	69
Answers to Practice Exercises	71

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.

- I. Overview (page 42)**
- II. Special case—Simple analysis (pages 43–46)**
- III. Assessing multiplicative interaction (pages 46–52)**
- IV. The E, V, W model—A general model containing a (0, 1) exposure and potential confounders and effect modifiers (pages 52–61)**
- V. Logistic model for matched data (pages 61–64)**

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 which 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 which 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.
6. Given a study situation involving matched case-control data:
 - a. state or recognize a logistic model for the analysis of matched data that controls for both matched and unmatched variables;
 - b. state how matching is incorporated into a logistic model using dummy variables;
 - c. state or recognize the expression for the odds ratio for the exposure–disease effect that controls for both matched and unmatched variables;
 - d. state or recognize null hypotheses for no interaction effect, or for no exposure–disease effect, given no interaction effect.

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
- Matched data

General logistic model formula :

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

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

α, β_i = unknown parameters

D = dichotomous outcome

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; controlling for several confounders and interaction terms; and analysis of matched data. In each case, we consider the definitions of variables in the model and the formula for the odds ratio describing the exposure–disease relationship.

Recall that the general logistic model for k independent variables may be written as $P(\mathbf{X})$ equals 1 over 1 plus e to minus the quantity α plus the sum of $\beta_i X_i$, where $P(\mathbf{X})$ 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 which we need to estimate from data obtained for a group of subjects on the X 's and on D , a dichotomous disease outcome variable.

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

$$\begin{aligned} \text{ROR} &= e^{\sum_{i=1}^k \beta_i (X_{1i} - X_{0i})} \\ &= \prod_{i=1}^k e^{\beta_i (X_{1i} - X_{0i})} \end{aligned}$$

\mathbf{X}_1 specification of \mathbf{X} for subject 1

\mathbf{X}_0 specification of \mathbf{X} for subject 0

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.

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, \dots, 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

$X_1 = E = \text{exposure } (0, 1)$

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

	E	E
D	a	b
D	c	d

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**.

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.

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

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 $P(\mathbf{X})$ 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 Chapter 3.

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

The logit form of the logistic model we have just defined is of the form $\text{logit } P(\mathbf{X})$ 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.

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

$$E = 1: R_1 = \Pr(D = 1 | E = 1)$$

$$E = 0: R_0 = \Pr(D = 1 | E = 0)$$

The term $P(\mathbf{X})$ 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 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 R_0 , which is the conditional probability that D equals 1 given that E equals 0.

44 2. Important Special Cases of the Logistic Model

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

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

into ROR formula:

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

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

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

algebra

$$= \left(e^{\beta_1} \right)$$

General ROR formula used for other special cases

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 .

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.

For E equal to 1, we can write \mathbf{R}_1 by substituting the value E equals 1 into the model formula for $P(\mathbf{X})$. 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 .

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 .

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.

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 :

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

Simple analysis :

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

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

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

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

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

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

SIMPLE ANALYSIS SUMMARY

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

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

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

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 i th X variable for group 1 persons and X_{0i} denotes the value of the i th 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.

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 X 's denoted by the **bold 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 X 's denoted by the **bold X** is written as \mathbf{X}_0 and equals the collection of one value X_{01} , which equals 0.

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 which 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.

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 .

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

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

Simple analysis: does not need computer

Other special cases: require computer

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.

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

$$X_1 = A = (0, 1) \text{ variable}$$

$$X_2 = B = (0, 1) \text{ variable}$$

Interaction: equation involving RORs for combinations of A and B

$$R_{AB} = \text{risk given } A, B \\ = \Pr(D = 1 | A, B)$$

$B=1$ $B=0$

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

Note: above table not for simple analysis.

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

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.

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 .

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} .

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

		$B=1$	$B=0$	
		$A=1$		
		$A=0$		
				referent cell

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

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

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

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

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

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

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

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

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

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.

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.

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.

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 minus R_{AB} , where A takes on the values 0 and 1 and B takes on the values 0 and 1.

DEFINITION

$OR_{11} = OR_{10} \times OR_{01}$

↑ ↑
 no interaction multiplication
 on a on a
 multiplicative multiplicative
 scale 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 :

$$\begin{pmatrix} \text{effect of } A \text{ and } B \\ \text{acting together} \end{pmatrix} = \begin{pmatrix} \text{combined effect} \\ \text{of } A \text{ and } B \\ \text{acting separately} \end{pmatrix}$$

$$\begin{matrix} \uparrow & \uparrow \\ OR_{11} & OR_{10} \times OR_{01} \\ & \text{multiplicative} \\ & \text{scale} \end{matrix}$$

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 :

$$OR_{11} = OR_{10} \times OR_{01}$$

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

EXAMPLE

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

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

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

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

$$\text{OR}_{11} \stackrel{?}{=} \text{OR}_{10} \times \text{OR}_{01}$$

$$14.4 \stackrel{?}{=} 7.2 \times 2.0$$

↑
14.4

Yes

$$B=1 \qquad \qquad B=0$$

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

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

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

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

$$\text{OR}_{11} \stackrel{?}{=} \text{OR}_{10} \times \text{OR}_{01}$$

$$30.0 \stackrel{?}{=} 7.2 \times 2.0$$

↑

No

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 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, Chapter 19

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

Logistic model variables:

$$\begin{aligned} X_1 &= A(0, 1) \\ X_2 &= B(0, 1) \\ X_3 &= A \times B \quad \text{interaction effect variable} \end{aligned}$$

main effects

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

where

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

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

We now define a logistic model which 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.

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

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

$$\Leftrightarrow H_0 : OR_{11} = OR_{10} \times OR_{01}$$

$$\Leftrightarrow H_0 : \frac{OR_{11}}{OR_{10} \times OR_{01}} = 1$$

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

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

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

$$H_0 : \text{no interaction} \Leftrightarrow \beta_3 = 0$$

Test result *Model*

$$\text{not significant} \Rightarrow \alpha + \beta_1 A + \beta_2 B$$

$$\text{significant} \Rightarrow \alpha + \beta_1 A + \beta_2 B + \beta_3 AB$$

MAIN POINT:

Interaction test \Rightarrow test for product terms

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.

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 $\text{logit } P(\mathbf{X})$ 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.

A description of methods for testing hypotheses for logistic regression models is beyond the scope of this presentation (see Chapter 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)

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

H_0 : no interaction (multiplicative)

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

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 $\text{logit } P(\mathbf{X})$ 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, \dots, 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.

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

control variables

$$\left\{ \begin{array}{l} D = \text{CHD}_{(0, 1)} \\ E = \text{CAT}_{(0, 1)} \\ C_1 = \text{AGE}_{\text{continuous}} \\ C_2 = \text{CHL}_{\text{continuous}} \\ C_3 = \text{SMK}_{(0, 1)} \\ C_4 = \text{ECG}_{(0, 1)} \\ C_5 = \text{HPT}_{(0, 1)} \end{array} \right.$$

EXAMPLE (continued)

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:

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

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

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

Parameters:

α , β , γ 's, and δ 's instead of α and β 's

where

β : exposure variable

γ 's: potential confounders

δ 's: potential interaction variables

Corresponding to these variables is 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.

The model is written 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 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.

Note that the parameters in this model are denoted as α , β , γ 's, and δ 's, whereas previously we denoted all parameters other than the constant α as β_i 's. We use β , γ 's, and δ 's here to distinguish different types of variables in the model. The parameter β indicates the coefficient of the exposure variable, the γ 's indicate the coefficients of the potential confounders in the model, and the δ 's 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, \dots, 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

$k = p_1 + p_2 + 1 = \# \text{ of variables in model}$
 $p_1 = \# \text{ of potential confounders}$
 $p_2 = \# \text{ of potential interactions}$
 $1 = \text{exposure variable}$

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 the 1 denotes the exposure variable.

CHD EXAMPLE

$p_1 = 5: \text{AGE, CHL, SMK, ECG, HPT}$
 $p_2 = 2: \text{CAT} \times \text{CHL, CAT} \times \text{HPT}$
 $p_1 + p_2 + 1 = 5 + 2 + 1 = 8$

- V_1, \dots, V_{p_1} are potential confounders
- V 's are functions of C 's

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

CHD EXAMPLE

$V_1 = \text{AGE}, V_2 = \text{CHL}, V_3 = \text{SMK},$
 $V_4 = \text{ECG}, V_5 = \text{HPT}$

- W_1, \dots, W_{p_2} are potential effect modifiers
- W 's are functions of C 's

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

In the CHD study example above, there are p_1 equal to five potential confounders, namely, the five control variables, and there are p_2 equal to two interaction variables, the first of which is $\text{CAT} \times \text{CHL}$ and the second is $\text{CAT} \times \text{HPT}$. 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.

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**.

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

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

Following the V 's, we define p_2 variables which 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 .

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

The CHD example above has two W 's, namely, CHL and HPT , that go into the model as product terms of the form $\text{CAT} \times \text{CHL}$ and $\text{CAT} \times \text{HPT}$.

CHD EXAMPLE

$W_1 = \text{CHL}, W_2 = \text{HPT}$

REFERENCES FOR CHOICE OF V's AND W's FROM C's

- Chapter 6: Modeling Strategy Guidelines
- *Epidemiologic Research*, Chapter 21

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

EXAMPLE

$$\begin{aligned}C_1 &= \text{AGE}, C_2 = \text{RACE}, C_3 = \text{SEX} \\V_1 &= \text{AGE}, V_2 = \text{RACE}, V_3 = \text{SEX} \\W_1 &= \text{AGE}, W_2 = \text{SEX} \\p_1 &= 3, p_2 = 2, k = p_1 + p_2 + 1 = 6\end{aligned}$$

NOTE

W's ARE SUBSET OF V's

EXAMPLE

$$\begin{array}{ll}\cancel{V_1 = \text{AGE}}, & \cancel{V_2 = \text{RACE}} \\ \cancel{W_1 = \text{AGE}}, & \cancel{W_2 = \text{SEX}}\end{array}$$

$$\begin{aligned}\text{logit } P(\mathbf{X}) &= \alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 + \cdots + \gamma_{p_1} V_{p_1} \\&\quad + \delta_1 E W_1 + \delta_2 E W_2 + \cdots + \delta_{p_2} E W_{p_2}\end{aligned}$$

where

- β = coefficient of E
- γ 's = coefficient of V 's
- δ 's = coefficient of W 's

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

$$+ \sum_{i=1}^{p_1} \gamma_i V_i + E \sum_{j=1}^{p_2} \delta_j W_j$$

It is beyond the scope of this presentation 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 (Chapter 6) on modeling strategies and in Chapter 21 of *Epidemiologic Research* by Kleinbaum, Kupper, and Morgenstern.

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 that although more details are given in the above references, 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.

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.

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

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, \dots, C_p fixed

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

- γ_i terms not in 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 Chapter 3 on coding)

Interaction :

$$\text{ROR} = \exp\left(\beta + \sum (\delta_j W_j)\right)$$

- $\delta_j \neq 0 \Rightarrow \text{OR depends on } W_j$
- Interaction \Rightarrow effect of E differs at different levels of W 's

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 versus 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.

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 .

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.

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 Chapter 3.

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 W 's.

- V 's not in OR formula but V 's in model, so OR formula controls confounding:

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

No interaction :

$$\text{all } \delta_j = 0 \Rightarrow \text{ROR} = \exp(\beta)$$

↑
constant

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \sum \gamma_i V_i$$

↑
confounding
effects adjusted

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.

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 V 's, 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 V 's.

EXAMPLE

The model :

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

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.

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

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

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.

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

$n = 609$ 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 \times CHL	$\hat{\delta}_1 = 0.0692$
CAT \times HPT	$\hat{\delta}_2 = -2.3318$

$$\widehat{\text{ROR}} = \exp(-12.6894 + 0.0692\text{CHL} - 2.3318\text{HPT})$$

exposure coefficient interaction coefficient

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 δ 's, 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.

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.

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 δ 's in the model.

EXAMPLE (continued)

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

 interaction variables

- CHL = 220, HPT = 1

$$\widehat{\text{ROR}} = \exp[-12.6894 + 0.0692(220) - 2.3318(1)] \\ = \exp(0.2028) = 1.22$$

- CHL = 200, HPT = 0

$$\widehat{\text{ROR}} = \exp[-12.6894 + 0.0692(200) - 2.3318(0)] \\ = \exp(1.1506) = 3.16$$

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

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

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

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 interaction variables in the model, and by interaction, we mean that the odds ratio changes as the values of the interaction variables change.

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.

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.

Thus, we see that depending on the values of the interaction variables, 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.

In general, when faced with an odds ratio expression involving interaction (*W*) variables, 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.

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

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

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

EXAMPLE (continued)

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

Fitted model:

<i>Variable</i>	<i>Coefficient</i>
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	$\hat{\gamma}_5 = 0.4392$

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

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.

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

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

	<i>interaction model</i>	<i>no interaction model</i>
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 Chapters 6 and 7).

V. Logistic Model for Matched Data (Chapter 8)

Focus: matched case-control studies

Principle:

matched analysis \Rightarrow stratified analysis

- strata are matched sets, e.g., pairs or combinations of matched sets, e.g., pooled pairs
- strata defined using dummy (indicator) variables

$E = (0, 1)$ exposure

C_1, C_2, \dots, C_p control variables

- some C 's matched by design
- remaining C 's not matched

We will now consider a special case of the logistic model and the corresponding odds ratio for the **analysis of matched data**. This topic is discussed in more detail in Chapter 8. Our focus here will be on matched case-control studies, although the formulae provided also apply to matched follow-up studies.

An important principle about modeling matched data is that a **matched analysis is a stratified analysis**. The strata are the matched sets, for example, the pairs in a matched pair design, or combinations of matched sets, such as pooled pairs within a close age range.

Moreover, when we use logistic regression to do a matched analysis, we define the strata using **dummy**, or indicator, variables.

In defining a model for a matched analysis, we again 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. We assume that some of these C variables have been matched in the study design, either by using pair matching, R -to-1 matching, or frequency matching. The remaining C variables have not been matched, but it is of interest to control for them, nevertheless.

62 2. Important Special Cases of the Logistic Model

$D = (0, 1)$ disease

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

Some X 's: V_{1i} dummy variables
(matched status)

Some X 's: V_{2i} variables
(potential confounders)

Some X 's: product terms EW_j
(potential interaction variables)

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \underbrace{\sum_{\text{matching}} \gamma_{1i} V_{1i}}_{\text{confounders}} + E \underbrace{\sum_{\text{interaction}} \delta_j W_j}_{\text{interaction}}$$

EXAMPLE

Pair matching by AGE, RACE, SEX

100 matched pairs

99 dummy variables

$$V_{1i} = \begin{cases} 1 & \text{if } i\text{th matched pair} \\ 0 & \text{otherwise} \end{cases}$$

$$i = 1, 2, \dots, 99$$

$$V_{11} = \begin{cases} 1 & \text{if first matched pair} \\ 0 & \text{otherwise} \end{cases}$$

$$V_{12} = \begin{cases} 1 & \text{if second matched pair} \\ 0 & \text{otherwise} \end{cases}$$

:

$$V_{1,99} = \begin{cases} 1 & \text{if 99th matched pair} \\ 0 & \text{otherwise} \end{cases}$$

Given the above context, we will 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 . We also have a collection of X 's that are dummy variables indicating the different matched strata; these variables are denoted as V_{1i} variables.

Further, we have a collection of X 's that are defined from the C 's not involved in the matching. These X 's represent potential confounders in addition to the matched variables and are denoted as V_{2i} variables.

Finally, we have a collection of X 's which are product terms of the form E times W_j , where the W 's denote potential effect modifiers. Note that the W 's will usually be defined in terms of the V_2 variables.

The logistic model for a matched analysis is shown here. Note that the γ_{1i} are coefficients of the dummy variables for the matching strata, the γ_{2i} are the coefficients of the potential confounders not involved in the matching, and the δ_j are the coefficients of the interaction variables.

As an example of dummy variables defined for matched strata, consider a study involving pair matching by age, race, and sex, and containing 100 matched pairs. Then the above model requires defining 99 dummy variables to incorporate the 100 matched pairs. For example, we can define these variables as V_{1i} equals 1 if an individual falls into the i th matched pair and 0 otherwise. Thus, 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 :

$$V_{11} = 1, V_{12} = V_{13} = \dots = V_{1, 99} = 0$$

99th matched set :

$$V_{1, 99} = 1, V_{11} = V_{12} = \dots = V_{1, 98} = 0$$

100th matched set :

$$V_{11} = V_{12} = \dots = V_{1, 99} = 0$$

Alternatively, when we use 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 last matched set will have all 99 dummy variables equal to 0.

$$\text{ROR} = \exp(\beta + \sum \delta_j W_j)$$

- OR formula for E, V, W model
- two types of V variables are controlled

EXAMPLE

Case-control study

2-to-1 matching

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

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

$$\begin{array}{l} C_1 = \text{AGE}, C_2 = \text{RACE}, C_3 = \text{SEX}, C_4 = \text{HOSPITAL} \\ \quad \quad \quad \underbrace{\qquad \qquad \qquad}_{\text{matched}} \\ C_5 = \text{SBP}, C_6 = \text{ECG} \\ \quad \quad \quad \underbrace{\qquad \qquad \qquad}_{\text{not matched}} \end{array}$$

$$n = 117 \text{ (39 matched sets)}$$

$$\begin{aligned} \text{logit } P(\mathbf{X}) = & \alpha + \beta \text{ SMK} + \sum_{i=1}^{38} \gamma_{1i} V_{1i} \\ & + \gamma_{21} \text{SBP} + \gamma_{22} \text{ECG} \\ & + \text{SMK}(\delta_1 \text{SBP} + \delta_2 \text{ECG}) \end{aligned}$$

The odds ratio formula for the matched analysis model is given by the expression ROR equals e to the quantity β plus the sum of the δ_j times the W_j . This is exactly the same odds ratio formula as given earlier for the E, V, W model for (0, 1) exposure variables. The matched analysis model is essentially an E, V, W model also, even though it contains two different types of V variables.

As an example of a matched pairs model, consider a case-control study using 2-to-1 matching that involves the following variables: The disease variable is myocardial infarction status (MI); the exposure variable is smoking status (SMK), a (0, 1) variable.

There are six C variables to be controlled. The first four of these variables—age, race, sex, and hospital status—are involved in the matching, and the last two variables—systolic blood pressure (SBP) and electrocardiogram status (ECG)—are not involved in the matching.

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.

A logistic model for the above situation is shown here. This model contains 38 terms of the form γ_{1i} times V_{1i} , where V_{1i} are dummy variables for the 39 matched sets. The model also contains two potential confounders, SBP and ECG, not involved in the matching, as well as two interaction variables involving these same two variables.

EXAMPLE (continued)

$$\text{ROR} = \exp(\beta + \delta_1 \text{SBP} + \delta_2 \text{ECG})$$

- does not contain V 's or γ 's
- V 's controlled as potential confounders

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. Note that this odds ratio expression does not contain any V terms or corresponding γ coefficients as such terms are potential confounders, not interaction variables. The V terms are nevertheless being controlled in the analysis because they are part of the logistic model being used.

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
- matched analyses

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 42)

A. Focus:

- simple analysis
- multiplicative interaction
- controlling several confounders and effect modifiers
- matched data

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

$$P(\mathbf{X}) = \frac{1}{1 + e^{-(\alpha + \sum_{i=1}^k \beta_i X_i)}}.$$

C. Logit form of logistic model:

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

D. General odds ratio formula:

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

II. Special case—Simple analysis (pages 43–46)

A. The model:

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

B. Logit form of the model:

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

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

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 46–52)

A. Definition of no interaction on a multiplicative scale:

$$\text{OR}_{11} = \text{OR}_{10} \times \text{OR}_{01},$$

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 52–61)

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

$$\text{ROR} = \exp(\beta).$$
- E. Examples of the E, V, W model: with interaction and without interaction

V. Logistic model for matched data (pages 61–64)

- A. Important principle about matching and modeling: A matched analysis is a stratified analysis. The strata are the matched sets, e.g., pairs in a matched pairs analysis. Must use dummy variables to distinguish among matching strata in the model.
- B. Specification of variables in the model:
 - i. Start with E and C_1, C_2, \dots, C_p .

- ii. Then specify a collection of V variables that are dummy variables indicating the matching strata, i.e., V_{1i} , where i ranges from 1 to $G-1$, if there are G strata.
- iii. Then specify V variables that correspond to potential confounders not involved in the matching (these are called V_{2i} variables).
- iv. Finally, specify a collection of W variables that correspond to potential effect modifiers not involved in the matching (these are called W_j variables).
- C. The logit form of a logistic model for matched data:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \sum_{i=1}^{G-1} \gamma_{1i} V_{1i} + \sum \gamma_{2i} V_{2i} + E \sum \delta_j W_j$$

- D. The odds ratio expression for the above matched analysis model:

$$\text{ROR}_{E=1 \text{ vs. } E=0} = \exp\left(\beta + \sum \delta_j W_j\right)$$
- E. The null hypothesis of no interaction in the above matched analysis model:

$$H_0: \text{ all } \delta_j = 0.$$
- F. Examples of a matched analysis model and odds ratio.

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

$$\text{logit } P(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$
- T F 8. An equation that describes “no interaction on a multiplicative scale” is given by

$$\text{OR}_{11} = \text{OR}_{10} / \text{OR}_{01}.$$
- T F 9. Given the model $\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma \text{SMK} + \delta E \times \text{SMK}$, 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

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 \text{AGE} + \gamma_2 \text{SBP} + \gamma_3 \text{CHL} + \delta_1 \text{AGE} \times \text{SBP} + \delta_2 \text{AGE} \times \text{CHL} + \delta_3 \text{SBP} \times \text{CHL}.$$
- T F 12. Given a logistic model of the form

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 \text{AGE} + \gamma_2 \text{SBP} + \gamma_3 \text{CHL},$$
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 $\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 \text{SMK} + \gamma_2 \text{SBP}$, 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

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 C_1 + \gamma_2 C_2 + E(\delta_1 C_1 + \delta_2 C_2).$$
- 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 $\exp(\beta + 20\delta_1 + 5\delta_2)$.

Given a matched pairs case-control study with 100 subjects (50 pairs), suppose that in addition to the variables involved in the matching, the variable physical activity level (PAL) was measured but not involved in the matching.

- T F 17. For the matched pairs study described above, assuming no pooling of matched pairs into larger strata, a logistic model for a matched analysis that contains an intercept term requires 49 dummy variables to distinguish among the 50 matched pair strata.
- T F 18. For the matched pairs study above, a logistic model assessing the effect of a (0, 1) exposure E and controlling for the confounding effects of the matched variables and the unmatched variable PAL plus the interaction effect of PAL with E is given by the expression

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \sum_{i=1}^{49} \gamma_i V_i + \gamma_{50} \text{PAL} + \delta E \times \text{PAL},$$

where the V_i are dummy variables that indicate the matched pair strata.

- T F 19. Given the model in Exercise 18, the odds ratio for the exposure-disease relationship that controls for matching and for the confounding and interactive effect of PAL is given by $\exp(\beta + \delta \text{PAL})$.
- T F 20. Again, given the model in Exercise 18, the null hypothesis for a test of no interaction on a multiplicative scale can be stated as $H_0: \delta = 0$.

Test

True or False (Circle T or F)

- T F 1. Given the simple analysis model, $\text{logit } P(\mathbf{X}) = \phi + \psi Q$, 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 $\text{logit } P(\mathbf{X}) = \alpha + \beta E$, 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 $\text{logit } P(\mathbf{X}) = \phi + \psi HPT + \rho ECG + \pi HPT \times ECG$, 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 $\text{logit } P(\mathbf{X}) = \alpha + \beta E + \phi \text{HPT} + \psi \text{ECG}$, 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

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 C_1 + \gamma_2 C_1^2 + \gamma_3 C_2 + \delta E C_1.$$
- T F 8. For the model in Question 7, if $C_1 = 5$ and $C_2 = 20$, then the odds ratio for the E , D relationship has the form $\exp(\beta + 20\delta)$.
- Given a 4-to-1 case-control study with 100 subjects (i.e., 20 matched sets), suppose that in addition to the variables involved in the matching, the variables obesity (OBS) and parity (PAR) were measured but not involved in the matching.
- T F 9. For the matched pairs study above, a logistic model assessing the effect of a $(0, 1)$ exposure E , controlling for the confounding effects of the matched variables and the unmatched variables OBS and PAR plus the interaction effects of OBS with E and PAR with E , is given by the expression

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \sum_{i=1}^{99} \gamma_{1i} V_{1i} + \gamma_{21} \text{OBS} + \gamma_{22} \text{PAR} + \delta_1 E \times \text{OBS} + \delta_2 E \times \text{PAR}.$$

- T F 10. Given the model in Question 9, the odds ratio for the exposure-disease relationship that controls for matching and for the confounding and interactive effects of OBS and PAR is given by $\exp(\beta + \delta_1 E \times \text{OBS} + \delta_2 E \times \text{PAR})$.

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 to 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.

11. 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.)

12. Using the model in Question 11, 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.

Suppose, instead of a follow-up study, that a matched pairs case-control study involving 200 pairs of bisexual males is performed, where the matching variables are NP and ASCM as described above, where PAR is also determined but is not involved in the matching, and where CON is the exposure variable.

13. Within the matched pairs case-control framework, state the logit form of a logistic model for assessing the effect of CON on HIV acquisition, controlling for PAR, NP, and ASCM as potential confounders and only PAR as an effect modifier.

14. Using the model in Question 13, give an expression for the risk of an exposed person ($CON = 1$) who is in the first matched pair and whose value for PAR is 1.

15. Using the model in Question 13, give an expression for the same odds ratio for the effect of CON, controlling for the confounding effects of NP, ASCM, and PAR and for the interaction effect of PAR.

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

72 2. Important Special Cases of the Logistic Model

8. F: $OR_{11} = OR_{10} \times OR_{01}$
9. T
10. F: $OR = \exp(\beta + \delta SMK)$
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
17. T
18. T
19. T
20. F: $H_0: \delta = 0$

3

Computing the Odds Ratio in Logistic Regression

Contents	
Introduction	74
Abbreviated Outline	74
Objectives	75
Presentation	76
Detailed Outline	92
Practice Exercises	95
Test	97
Answers to Practice Exercises	99

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.

- I. Overview (pages 76–77)
- II. Odds ratio for other codings of a dichotomous *E* (pages 77–79)
- III. Odds ratio for arbitrary coding of *E* (pages 79–82)
- IV. The model and odds ratio for a nominal exposure variable (no interaction case) (pages 82–84)
- V. The model and odds ratio for several exposure variables (no interaction case) (pages 85–87)
- VI. The model and odds ratio for several exposure variables with confounders and interaction (pages 87–91)

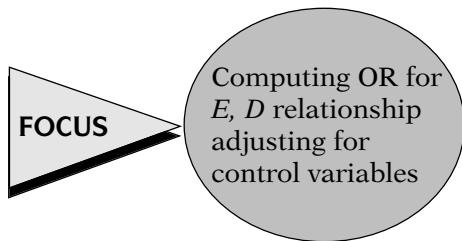
Objectives

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 which 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



- dichotomous E —arbitrary coding
- ordinal or interval E
- polytomous E
- several E 's

Chapter 2— E, V, W model:

- (0, 1) exposure
- confounders
- 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 $E \times W$)

The E, V, W model:

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

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.

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 which are ordinal or interval scaled variables. And, finally, we consider models involving several exposures, a special case of which involves a single polytomous exposure.

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.

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 **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 (γ 's) are coefficients of potential confounding variables denoted by the V 's, and the deltas (δ 's) are coefficients of potential interaction effects involving E separately with each of the W 's.

Adjusted odds ratio for effect of E adjusted for C's:

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

(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 C 's 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)

Focus: ✓ dichotomous
ordinal
interval

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

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

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.

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

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 .

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 .

EXAMPLES

$$(A) \quad a = 1, b = 0 \Rightarrow (a-b) = (1-0) = 1$$

$$\text{ROR} = \exp(1 \times \beta + 1 \times \sum \delta_j W_j)$$

$$(B) \quad a = 1, b = -1 \Rightarrow (a-b) = (1-[-1]) = 2$$

$$\text{ROR} = \exp(2\beta + 2 \sum \delta_j W_j)$$

$$(C) \quad a = 100, b = 0 \Rightarrow (a-b) = (100-0) = 100$$

$$\text{ROR} = \exp(100\beta + 100 \sum \delta_j W_j)$$

Coding	\widehat{ROR}
(A) $a = 1, b = 0$	$\widehat{ROR}_A = \exp\left(\hat{\beta}_A + \sum_{j=1}^{p_2} \hat{\delta}_{jA} W_j\right)$
(B) $a = 1, b = -1$	$\widehat{ROR}_B = \exp\left(2\hat{\beta}_B + \sum_{j=1}^{p_2} 2\hat{\delta}_{jB} W_j\right)$
(C) $a = 100, b = 0$	$\widehat{ROR}_C = \exp\left(100\hat{\beta}_C + \sum_{j=1}^{p_2} 100\hat{\delta}_{jC} W_j\right)$

same value
although
different codings
 different values
for different codings
 $\widehat{ROR}_A = \widehat{ROR}_B = \widehat{ROR}_C$
 $\hat{\beta}_A \neq \hat{\beta}_B \neq \hat{\beta}_C$
 $\hat{\delta}_{jA} \neq \hat{\delta}_{jB} \neq \hat{\delta}_{jC}$

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

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

CAT: (0, 1) versus other codings



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

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.

This model is written in **logit form** as $\text{logit } P(\mathbf{X})$ 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.

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.

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	$\hat{\gamma}_5 = 0.4392$

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

No interaction model: ROR fixed

$$(-1, 1) \text{ coding for CAT: } \hat{\beta} = 0.2989 = \left(\frac{0.5978}{2} \right)$$

$$\begin{aligned}\widehat{ROR} &= \exp(2\hat{\beta}) = \exp(2 \times 0.2989) \\ &= \exp(0.5978) \\ &= 1.82\end{aligned}$$

same \widehat{ROR} as for (0, 1) coding

$$\text{Note: } \widehat{ROR} \neq \exp(0.2989) = 1.35$$

↑
incorrect value

Fitting this no interaction model to the data, we obtain the estimates listed here.

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.

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:

dichotomous, ordinal or interval

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.

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

E^* (group 1) vs. E^{**} (group 2)

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

Same as

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

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.

The odds ratio formula for E^* versus 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 .

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.

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 $5\beta + 5 \sum \delta_j W_j$.

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 $2\beta + 2 \sum \delta_j W_j$.

EXAMPLE

$E = \text{SSU}$ = social support status (0–5)

(A) $\text{SSU}^* = 5$ vs. $\text{SSU}^{**} = 0$

$$\begin{aligned} \text{ROR}_{5,0} &= \exp [(\text{SSU}^* - \text{SSU}^{**}) \\ &\quad \beta + (\text{SSU}^* - \text{SSU}^{**}) \sum \delta_j W_j] \\ &= \exp [(5 - 0)\beta + (5 - 0) \sum \delta_j W_j] \\ &= \exp (5\beta + 5 \sum \delta_j W_j) \end{aligned}$$

(B) $\text{SSU}^* = 3$ vs. $\text{SSU}^{**} = 1$

$$\begin{aligned} \text{ROR}_{3,1} &= \exp [(3 - 1)\beta + (3 - 1) \sum \delta_j W_j] \\ &= \exp (2\beta + 2 \sum \delta_j W_j) \end{aligned}$$

EXAMPLE (continued)

$$(C) \text{ SSU}^* = 4 \text{ vs. } \text{SSU}^{**} = 2$$

$$\text{ROR}_{4, 2} = \exp[(4-2)\beta + (4-2)\sum \delta_j W_j]$$

$$= \exp(2\beta + 2\sum \delta_j W_j)$$

Note: ROR depends on the difference $(E^* - E^{**})$, e.g., $(3 - 1) = (4 - 2) = 2$

EXAMPLE

$E = \text{SBP}$ = systolic blood pressure (interval)

$$(A) \text{ SBP}^* = 160 \text{ vs. } \text{SBP}^{**} = 120$$

$$\text{ROR}_{160, 120} = \exp[(\text{SBP}^* - \text{SBP}^{**})\beta + (\text{SBP}^* - \text{SBP}^{**})\sum \delta_j W_j]$$

$$= \exp[(160 - 120)\beta + (160 - 120)\sum \delta_j W_j]$$

$$= \exp(40\beta + 40\sum \delta_j W_j)$$

$$(B) \text{ SBP}^* = 200 \text{ vs. } \text{SBP}^{**} = 120$$

$$\text{ROR}_{200, 120} = \exp[(200 - 120)\beta + (200 - 120)\sum \delta_j W_j]$$

$$= \exp(80\beta + 80\sum \delta_j W_j)$$

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^{**} .

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 δ_j times W_j .

No interaction:

$$\text{ROR}_{E^* \text{ vs. } E^{**}} = \exp[(E^* - E^{**})\beta]$$

If $(E^* - E^{**}) = 1$, then $\text{ROR} = \exp(\beta)$

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.

Thus, if E denotes SBP, then the quantity e to β gives the odds ratio for comparing any two groups which 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

$E = \text{SBP}$

$$\text{ROR} = \exp(\beta) \Rightarrow (\text{SBP}^* - \text{SBP}^{**}) = 1$$

not interesting \uparrow

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
----------------	-----	-----	-----	-----	-----

EXAMPLE (continued)

<u>SBP*</u>	<u>SBP**</u>	<u>OR</u>
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, \dots, E_q

- model
- odds ratio

Nominal variable: > 2 categories

e.g., ✓ occupational status in four groups

~~SSU (0 = 5) ordinal~~

k categories $\Rightarrow k - 1$ dummy variables

E_1, E_2, \dots, E_{k-1}

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 E 's, 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.

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.

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 $k = 4 \Rightarrow k - 1 = 3$
 $\quad \quad \quad \text{OCC}_1, \text{OCC}_2,$
 $\quad \quad \quad \text{OCC}_3$

where $\text{OCC}_i = \begin{cases} 1 & \text{if category } i \\ 0 & \text{if otherwise} \end{cases}$
for $i = 1, 2, 3$

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 i th occupational category and 0 otherwise, for i ranging from 1 to 3.

No interaction model:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta_1 E_1 + \beta_2 E_2 + \dots + \beta_{k-1} E_{k-1} + \sum_{i=1}^{p_1} \gamma_i V_i$$

$$\text{logit } P(\mathbf{X}) = \alpha + \beta_1 \text{OCC}_1 + \beta_2 \text{OCC}_2 + \beta_3 \text{OCC}_3 + \sum_{i=1}^{p_1} \gamma_i V_i$$

Specify \mathbf{E}^* and \mathbf{E}^{**} in terms of $k - 1$ dummy variables where

$$\mathbf{E} = (E_1, E_2, \dots, E_{k-1})$$

A no interaction model for a nominal exposure variable with k categories then takes the form $\text{logit } P(\mathbf{X})$ 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.

The corresponding model for four occupational status categories then becomes $\text{logit } P(\mathbf{X})$ equals α plus β_1 times OCC_1 plus β_2 times OCC_2 plus β_3 times OCC_3 plus the V terms.

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

$E = \text{occupational status (four categories)}$
 $\mathbf{E}^* = \text{category 3 vs. } \mathbf{E}^{**} = \text{category 1}$
 $\mathbf{E}^* = (\text{OCC}_1^* = 0, \text{OCC}_2^* = 0, \text{OCC}_3^* = 1)$
 $\mathbf{E}^{**} = (\text{OCC}_1^{**} = 1, \text{OCC}_2^{**} = 0, \text{OCC}_3^{**} = 0)$

Generally, define \mathbf{E}^* and \mathbf{E}^{**} as

$$\mathbf{E}^* = (E_1^*, E_2^*, \dots, E_{k-1}^*)$$

and

$$\mathbf{E}^{**} = (E_1^{**}, E_2^{**}, \dots, E_{k-1}^{**})$$

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 $\text{OCC}_3^{**} = 0$.

More generally, category \mathbf{E}^* is defined by the dummy variable values $E_1^*, E_2^*, \dots, E_{k-1}^*$, which are 0's or 1's. Similarly, category \mathbf{E}^{**} is defined by the values $E_1^{**}, E_2^{**}, \dots, E_{k-1}^{**}$, which is a different specification of 0's or 1's.

No interaction model

$$\text{ROR}_{E^* \text{ vs. } E^{**}}$$

$$= \exp [(E_1^* - E_1^{**})\beta_1 + (E_2^* - E_2^{**})\beta_2 + \dots + (E_{k-1}^* - E_{k-1}^{**})\beta_{k-1}]$$

EXAMPLE (OCC)

$$\begin{aligned}\text{ROR}_{3 \text{ vs. } 1} &= \exp \left[\begin{matrix} 0 & 1 \\ (OCC_1^* - OCC_1^{**})\beta_1 \\ 0 & 0 \\ (OCC_2^* - OCC_2^{**})\beta_2 \\ 1 & 0 \\ (OCC_3^* - OCC_3^{**})\beta_3 \end{matrix} \right] \\ &= \exp [(0 - 1)\beta_1 + (0 - 0)\beta_2 + (1 - 0)\beta_3] \\ &= \exp [(-1)\beta_1 + (0)\beta_2 + (1)\beta_3] \\ &= \exp (-\beta_1 + \beta_3)\end{aligned}$$

$$\widehat{\text{ROR}} = \exp (-\hat{\beta}_1 + \hat{\beta}_3)$$

E^* = category 3 vs. E^{**} = category 2:

$$E^* = (OCC_1^* = 0, OCC_2^* = 0, OCC_3^* = 1)$$

$$E^{**} = (OCC_1^{**} = 0, OCC_2^{**} = 1, OCC_3^{**} = 0)$$

$$\begin{aligned}\text{ROR}_{3 \text{ vs. } 2} &= \exp [(0 - 0)\beta_1 + (0 - 1)\beta_2 + (1 - 0)\beta_3] \\ &= \exp [(0)\beta_1 + (-1)\beta_2 + (1)\beta_3] \\ &= \exp (-\beta_2 + \beta_3)\end{aligned}$$

Note: $\text{ROR}_{3 \text{ vs. } 1} = \exp (-\beta_1 + \beta_3)$

The **general odds ratio formula** for comparing two categories, E^* versus 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.

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 .

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 $(-\beta_1) + \beta_3$.

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 $(-\hat{\beta}_1) + \hat{\beta}_3$.

In contrast, if category 3 is compared to category 2, then E^* takes on the values 0, 0, and 1 as before, whereas E^{**} is now defined by $OCC_1^{**} = 0$, $OCC_2^{**} = 1$, and $OCC_3^{**} = 0$.

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 $(-\beta_2) + \beta_3$.

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, \dots, 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 E 's can be any kind of variable—dichotomous, ordinal, or interval.

EXAMPLE

$$\begin{aligned}E_1 &= \text{SMK (0,1)} \\E_2 &= \text{PAL (ordinal)} \\E_3 &= \text{SBP (interval)}\end{aligned}$$

No interaction model:

$$\begin{aligned}\text{logit } P(\mathbf{X}) &= \alpha + \beta_1 E_1 + \beta_2 E_2 + \dots + \beta_q E_q \\&\quad + \sum_{i=1}^{p_1} \gamma_i V_i\end{aligned}$$

- $q \neq k - 1$ in general

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).

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 E 's 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.

EXAMPLE

$$\begin{aligned}\text{logit } P(\mathbf{X}) &= \alpha + \beta_1 \text{SMK} + \beta_2 \text{PAL} + \beta_3 \text{SBP} \\&\quad + \sum_{i=1}^{p_1} \gamma_i V_i\end{aligned}$$

\mathbf{E}^* vs. \mathbf{E}^{**}

$$\begin{aligned}\mathbf{E}^* &= (E_1^*, E_2^*, \dots, E_q^*) \\ \mathbf{E}^{**} &= (E_1^{**}, E_2^{**}, \dots, E_q^{**})\end{aligned}$$

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^{**} .

$$\begin{aligned}\text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} &= \exp \left[(E_1^* - E_1^{**}) \beta_1 \right. \\&\quad \left. + (E_2^* - E_2^{**}) \beta_2 + \dots \right. \\&\quad \left. + (E_q^* - E_q^{**}) \beta_q \right]\end{aligned}$$

The general odds ratio formula for comparing \mathbf{E}^* versus \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 .

In general

- q variables $\neq k - 1$ dummy variables

EXAMPLE

$$\text{logit } P(X) = \alpha + \beta_1 \text{SMK} + \beta_2 \text{PAL} \\ + \beta_3 \text{SBP} \\ + \gamma_1 \text{AGE} + \gamma_2 \text{SEX}$$

Nonsmoker, PAL = 25, SBP = 160
vs.

Smoker, PAL = 10, SBP = 120

$$\mathbf{E}^* = (\text{SMK}^*=0, \text{PAL}^*=25, \text{SBP}^*=160)$$

$$\mathbf{E}^{**} = (\text{SMK}^{**}=1, \text{PAL}^{**}=10, \text{SBP}^{**}=120)$$

AGE and SEX fixed, but unspecified

$$\text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} = \exp[(\text{SMK}^* - \text{SMK}^{**})\beta_1 \\ + (\text{PAL}^* - \text{PAL}^{**})\beta_2 \\ + (\text{SBP}^* - \text{SBP}^{**})\beta_3]$$

$$= \exp[(0 - 1)\beta_1 + (25 - 10)\beta_2 + (160 - 120)\beta_3]$$

$$= \exp[(-1)\beta_1 + (15)\beta_2 + (40)\beta_3]$$

$$= \exp(-\beta_1 + 15\beta_2 + 40\beta_3)$$

$$\widehat{\text{ROR}} = \exp(-\hat{\beta}_1 + 15\hat{\beta}_2 + 40\hat{\beta}_3)$$

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.

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.

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 $\text{SBP}^{**}=120$.

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.

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 ,

which equals e to $(0 - 1)$ times β_1 plus $(25 - 10)$ times β_2 plus $(160 - 120)$ times β_3 ,

which equals e to the quantity -1 times β_1 plus 15 times β_2 plus 40 times β_3 ,

which reduces to e to the quantity $-\beta_1$ plus $15\beta_2$ plus $40\beta_3$.

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

$$\begin{aligned}\text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} &= \exp[(1-1)\beta_1 + (25-5)\beta_2 + (160-200)\beta_3] \\ &= \exp[(0)\beta_1 + (20)\beta_2 + (-40)\beta_3] \\ &= \exp(20\beta_2 - 40\beta_3)\end{aligned}$$

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

EXAMPLE: The Variables

$$E_1 = \text{SMK}, E_2 = \text{PAL}, E_3 = \text{SBP}$$

$$V_1 = \text{AGE} = W_1, V_2 = \text{SEX} = W_2$$

$$E_1W_1 = \text{SMK} \times \text{AGE}, \quad E_1W_2 = \text{SMK} \times \text{SEX}$$

$$E_2W_1 = \text{PAL} \times \text{AGE}, \quad E_2W_2 = \text{PAL} \times \text{SEX}$$

$$E_3W_1 = \text{SBP} \times \text{AGE}, \quad E_3W_2 = \text{SBP} \times \text{SEX}$$

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.

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

$$\begin{aligned}\text{logit } P(\mathbf{X}) &= \alpha + \beta_1 \text{SMK} + \beta_2 \text{PAL} + \beta_3 \text{SBP} \\ &\quad + \gamma_1 \text{AGE} + \gamma_2 \text{SEX} \\ &\quad + \text{SMK}(\delta_{11} \text{AGE} + \delta_{12} \text{SEX}) \\ &\quad + \text{PAL}(\delta_{21} \text{AGE} + \delta_{22} \text{SEX}) \\ &\quad + \text{SBP}(\delta_{31} \text{AGE} + \delta_{32} \text{SEX})\end{aligned}$$

The corresponding model is given by logit $P(\mathbf{X})$ 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 δ 's 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

$$\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}$$

$$\mathbf{E}^* = (\text{SMK}^*=0, \text{PAL}^*=25, \text{SBP}^*=160)$$

$$\mathbf{E}^{**} = (\text{SMK}^{**}=1, \text{PAL}^{**}=10, \text{SBP}^{**}=120)$$

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): β 's only

ROR (interaction): β 's and δ 's

EXAMPLE (continued)

$$\begin{aligned} \text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} &= \exp[(\text{SMK}^* - \text{SMK}^{**})\beta_1 \\ &\quad + (\text{PAL}^* - \text{PAL}^{**})\beta_2 \\ &\quad + (\text{SBP}^* - \text{SBP}^{**})\beta_3 \\ &\quad + \delta_{11}(\text{SMK}^* - \text{SMK}^{**})\text{AGE} \\ &\quad + \delta_{12}(\text{SMK}^* - \text{SMK}^{**})\text{SEX} \\ &\quad + \delta_{21}(\text{PAL}^* - \text{PAL}^{**})\text{AGE} \\ &\quad + \delta_{22}(\text{PAL}^* - \text{PAL}^{**})\text{SEX} \\ &\quad + \delta_{31}(\text{SBP}^* - \text{SBP}^{**})\text{AGE} \end{aligned}$$

$$\begin{aligned} \text{ROR} &= \exp [(0 - 1)\beta_1 + (25 - 10)\beta_2 \\ &\quad + (160 - 120)\beta_3 \end{aligned}$$

interaction with SMK

$$\begin{aligned} &+ \delta_{11}(0 - 1)\text{AGE} + \delta_{12}(0 - 1)\text{SEX} \\ \text{interaction with PAL} \\ &+ \delta_{21}(25 - 10)\text{AGE} + \delta_{22}(25 - 10)\text{SEX} \\ \text{interaction with SBP} \\ &+ \delta_{31}(160 - 120)\text{AGE} + \delta_{32}(160 - 120)\text{SEX} \end{aligned}$$

$$\begin{aligned} &= \exp(-\beta_1 + 15\beta_2 + 40\beta_3 \\ &\quad - \delta_{11}\text{AGE} - \delta_{12}\text{SEX} \\ &\quad + 15\delta_{21}\text{AGE} + 15\delta_{22}\text{SEX} \\ &\quad + 40\delta_{31}\text{AGE} + 40\delta_{32}\text{SEX}) \end{aligned}$$

$$\begin{aligned} &= \exp(-\beta_1 + 15\beta_2 + 40\beta_3 \\ &\quad + \text{AGE}(-\delta_{11} + 15\delta_{21} + 40\delta_{31}) \\ &\quad + \text{SEX}(-\delta_{12} + 15\delta_{22} + 40\delta_{32})) \end{aligned}$$

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.

The odds ratio formula for our example then becomes 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 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 $(\text{SMK}^* - \text{SMK}^{**})$ times AGE, and the second of these terms is δ_{12} times the difference $(\text{SMK}^* - \text{SMK}^{**})$ times 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 δ 's.

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.

After subtraction, this expression reduces to the expression shown here at the left.

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:

$$\widehat{\text{ROR}} = \exp\left[-\hat{\beta}_1 + 15\hat{\beta}_2 + 40\hat{\beta}_3 + 35(-\hat{\delta}_{11} + 15\hat{\delta}_{21} + 40\hat{\delta}_{31}) + 1(-\hat{\delta}_{12} + 15\hat{\delta}_{22} + 40\hat{\delta}_{32})\right]$$

$$\begin{aligned}\widehat{\text{ROR}} = & \exp\left(-\hat{\beta}_1 + 15\hat{\beta}_2 + 40\hat{\beta}_3\right. \\ & - 35\hat{\delta}_{11} + 525\hat{\delta}_{21} + 1400\hat{\delta}_{31} \\ & \left.- \hat{\delta}_{12} + 15\hat{\delta}_{22} + 40\hat{\delta}_{32}\right)\end{aligned}$$

General model

Several exposures

Confounders

Effect modifiers

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 1400 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.

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.

$$\begin{aligned}\text{logit } P(\mathbf{X}) = & \alpha + \beta_1 E_1 + \beta_2 E_2 + \dots + \beta_q E_q \\ & + \sum_{i=1}^{p_1} \gamma_i V_i \\ & + E_1 \sum_{j=1}^{p_2} \delta_{1j} W_j + E_2 \sum_{j=1}^{p_2} \delta_{2j} W_j \\ & + \dots + E_q \sum_{j=1}^{p_2} \delta_{qj} W_j\end{aligned}$$

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 j th 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 j th 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:

$$\mathbf{E}^* = (E_1^*, E_2^*, \dots, E_q^*)$$

vs.

$$\mathbf{E}^{**} = (E_1^{**}, E_2^{**}, \dots, E_q^{**})$$

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 $E_1^*, E_2^*, \dots, E_q^*$; group \mathbf{E}^{**} is specified by a different collection of values $E_1^{**}, E_2^{**}, \dots, E_q^{**}$.

General Odds Ratio Formula:

$$\begin{aligned} \text{ROR}_{E^* \text{ vs. } E^{**}} &= \exp \left[(E_1^* - E_1^{**})\beta_1 \right. \\ &\quad + (E_2^* - E_2^{**})\beta_2 + \dots + (E_q^* - E_q^{**})\beta_q \\ &\quad + (E_1^* - E_1^{**}) \sum_{j=1}^{p_2} \delta_j W_j \\ &\quad + (E_2^* - E_2^{**}) \sum_{j=1}^{p_2} \delta_{2j} W_j \\ &\quad \left. + \dots + (E_q^* - E_q^{**}) \sum_{j=1}^{p_2} \delta_{qj} W_j \right] \end{aligned}$$

The general odds ratio formula for comparing two such specifications, \mathbf{E}^* versus \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 $(E^* - E^{**})$ times the sum of δ times W , where each of these latter terms correspond to interactions involving a different exposure variable.

EXAMPLE: $q = 3$

$$\begin{aligned} \text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} &= \exp[(\text{SMK} * -\text{SMK}^{**})\beta_1 \\ &\quad + (\text{PAL} * -\text{PAL}^{**})\beta_2 + (\text{SBP} * -\text{SBP}^{**})\beta_3 \\ &\quad + \delta_{11}(\text{SMK} * -\text{SMK}^{**})\text{AGE} \\ &\quad + \delta_{12}(\text{SMK} * -\text{SMK}^{**})\text{SEX} \\ &\quad + \delta_{21}(\text{PAL} * -\text{PAL}^{**})\text{AGE} \\ &\quad + \delta_{22}(\text{PAL} * -\text{PAL}^{**})\text{SEX} \\ &\quad + \delta_{31}(\text{SBP} * -\text{SBP}^{**})\text{AGE} \\ &\quad + \delta_{32}(\text{SBP} * -\text{SBP}^{**})\text{SEX}] \end{aligned}$$

- AGE and SEX controlled as V 's as well as W 's
- ROR's depend on values of W 's (AGE and SEX)

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.

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 (Chapter 4), we consider how the method of maximum likelihood is used to estimate the parameters of the logistic model. And in Chapter 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: $\text{logit } P(\mathbf{X}) = \alpha + \beta E + \sum_{i=1}^{p_1} \gamma_i V_i + \sum_{j=1}^{p_2} \delta_j W_j$.
 - 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: $a = 1, b = 0: \text{ROR} = \exp(\beta)$
 $a = 1, b = -1: \text{ROR} = \exp(2\beta)$
 $a = 100, b = 0: \text{ROR} = \exp(100\beta)$
- 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 E^* and E^{**} are any two values of E to be compared, the odds ratio formula becomes
- $$\text{ROR}_{E^* \text{ vs. } E^{**}} = \exp\left[(E^* - E^{**})\beta + (E^* - E^{**}) \sum_{j=1}^{p_2} \delta_j W_j\right]$$
- B. Examples: $E = \text{SSU} = \text{social support status (0–5)}$
 $E = \text{SBP} = \text{systolic blood pressure (interval)}$.
- C. No interaction odds ratio formula:

$$\text{ROR}_{E^* \text{ vs. } E^{**}} = \exp[(E^* - E^{**})\beta].$$
- 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, \dots, 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 $\text{OCC}_1, \text{OCC}_2$, 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:

$$\text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} = \exp \left[\begin{aligned} & \left(E_1^* - E_1^{**} \right) \beta_1 + \left(E_2^* - E_2^{**} \right) \beta_2 \\ & + \cdots + \left(E_{k-1}^* - E_{k-1}^{**} \right) \beta_{k-1} \end{aligned} \right]$$

where $\mathbf{E}^* = (E_1^*, E_2^*, \dots, E_{k-1}^*)$ and $\mathbf{E}^{**} = (E_1^{**}, E_2^{**}, \dots, E_{k-1}^{**})$ 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:

$$\text{ROR}_{\text{OCC}^* \text{ vs. } \text{OCC}^{**}} = \exp \left[\begin{aligned} & \left(\text{OCC}_1^* - \text{OCC}_1^{**} \right) \beta_1 + \left(\text{OCC}_2^* - \text{OCC}_2^{**} \right) \beta_2 \\ & + \left(\text{OCC}_3^* - \text{OCC}_3^{**} \right) \beta_3 \end{aligned} \right].$$

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, \dots, 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:

$$\text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} = \exp[(E_1^* - E_1^{**})\beta_1 + (E_2^* - E_2^{**})\beta_2 + \dots + (E_q^* - E_q^{**})\beta_q]$$

where $\mathbf{E}^* = (E_1^*, E_2^*, \dots, E_q^*)$ and $\mathbf{E}^{**} = (E_1^{**}, E_2^{**}, \dots, E_q^{**})$ are two specifications of the collection of exposure variables to be compared.

- D. Example of odds ratio involving three exposure variables:

$$\begin{aligned} \text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} = \exp[& (\text{SMK}^* - \text{SMK}^{**})\beta_1 + (\text{PAL}^* - \text{PAL}^{**})\beta_2 \\ & + (\text{SBP}^* - \text{SBP}^{**})\beta_3]. \end{aligned}$$

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:

$$\begin{aligned} \text{logit P}(\mathbf{X}) = & \alpha + \beta_1 E_1 + \beta_2 E_2 + \dots + \beta_q E_q + \sum_{i=1}^{p_1} \gamma_i V_i + E_1 \sum_{j=1}^{p_2} \delta_{1j} W_j \\ & + E_2 \sum_{j=1}^{p_2} \delta_{2j} W_j + \dots + E_q \sum_{j=1}^{p_2} \delta_{qj} W_j \end{aligned}$$

- D. The general odds ratio formula:

$$\begin{aligned} \text{ROR}_{\mathbf{E}^* \text{ vs. } \mathbf{E}^{**}} = \exp[& (E_1^* - E_1^{**})\beta_1 + (E_2^* - E_2^{**})\beta_2 + \dots + (E_q^* - E_q^{**})\beta_q \\ & + (E_1^* - E_1^{**}) \sum_{j=1}^{p_2} \delta_{1j} W_j + (E_2^* - E_2^{**}) \sum_{j=1}^{p_2} \delta_{2j} W_j \\ & + \dots + (E_q^* - E_q^{**}) \sum_{j=1}^{p_2} \delta_{qj} W_j] \end{aligned}$$

Practice Exercises

Given the model

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1(\text{SMK}) + \gamma_2(\text{HPT}) + \delta_1(E \times \text{SMK}) + \delta_2(E \times \text{HPT}),$$

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 $\text{CHL}=200$ to a person with $\text{CHL}=140$ 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

$$\text{logit } P(\mathbf{X}) = \alpha + \beta_1(\text{SMK}) + \beta_2(\text{ASB}) + \gamma_1(\text{AGE}) \\ + \delta_1(\text{SMK} \times \text{AGE}) + \delta_2(\text{ASB} \times \text{AGE}),$$

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
 $\text{logit } P(\mathbf{X}) = \alpha + \beta \text{CAT} + \gamma_1 \text{AGE} + \gamma_2 \text{CHL}$,
 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 β 's that are computed from the three different coding schemes?
2. Suppose the model in Question 1 is revised as follows:
 $\text{logit } P(\mathbf{X}) = \alpha + \beta \text{CAT} + \gamma_1 \text{AGE} + \gamma_2 \text{CHL} + \text{CAT}(\delta_1 \text{AGE} + \delta_2 \text{CHL})$.

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
 $\text{logit } P(\mathbf{X}) = \alpha + \beta \text{SSU} + \gamma_1 \text{AGE} + \gamma_2 \text{SEX} + \text{SSU}(\delta_1 \text{AGE} + \delta_2 \text{SEX})$,
 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. $\text{logit } P(\mathbf{X}) = \alpha + \beta_1 R_1 + \beta_2 R_2 + \gamma_1 \text{AGE} + \gamma_2 \text{SMK} + \gamma_3 \text{RACE} + \gamma_4 \text{SEX} + R_1(\delta_{11} \text{AGE} + \delta_{12} \text{SMK} + \delta_{13} \text{RACE} + \delta_{14} \text{SEX}) + R_2(\delta_{21} \text{AGE} + \delta_{22} \text{SMK} + \delta_{23} \text{RACE} + \delta_{24} \text{SEX}).$
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	
Introduction	102
Abbreviated Outline	102
Objectives	103
Presentation	104
Detailed Outline	120
Practice Exercises	121
Test	122
Answers to Practice Exercises	124

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, called 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 104)**
- II. Background about maximum likelihood procedure (pages 104–105)**
- III. Unconditional versus conditional methods (pages 105–109)**
- IV. The likelihood function and its use in the ML procedure (pages 109–115)**
- V. Overview on statistical inferences for logistic regression (pages 115–119)**

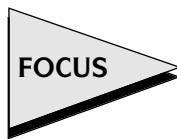
Objectives

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

1. State or recognize when to use unconditional versus conditional ML methods.
2. State or recognize what is a likelihood function.
3. State or recognize that the likelihood functions for unconditional versus conditional ML methods are different.
4. State or recognize that unconditional versus 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



- 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

- $\text{ML} = \text{LS}$ 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 to 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 Versus Conditional Methods

Two alternative ML approaches

- (1) unconditional method
 - (2) conditional method
- require different computer programs
 - user must choose appropriate program

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 programs. Thus, researchers using logistic regression modeling must decide which of these two programs 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 number of parameters is *small* relative to number of subjects

Conditional—preferred if number of parameters is *large* relative to number of subjects

Small vs. large? debatable

Guidelines provided here

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.

Exactly what is small versus 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 = (12)
(including intercept)

small relative to n = (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 i th 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)

Case-control study

100 matched pairs

Logistic model for matching:

- uses dummy variables for matching strata
- 99 dummy variables for 100 strata
- E , SMK, and $E \times$ SMK also in model

Number of parameters =

$$\begin{array}{ccccccc} 1 & + & 99 & + & 3 & = & (103) \\ \uparrow & & \uparrow & & \uparrow & & \\ \text{dummy} & & & & & & \\ \text{intercept} & \text{variables} & E, \text{SMK} & E \times \text{SMK} & & & \end{array}$$

large relative to 100 matched

pairs $\Rightarrow (n = 200)$

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.

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 covariate SMK, and perhaps even an interaction term of the form $E \times$ SMK.

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 $E \times$ 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 8: Analysis of Matched Data Using Logistic Regression

A detailed discussion of logistic regression for matched data is provided in Chapter 8.

Guidelines

- use *conditional* if matching
- use *unconditional* if no matching and number of variables not too large

The above examples indicate the following guidelines regarding the choice between unconditional and conditional ML methods or programs:

- 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 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 to 15 confounders
- 10 to 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 to 15 confounders and 10 to 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.

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: Conditional Required

Pair-matched case-control study
measure of effect: OR

EXAMPLE: (Continued)

Assume: only variables controlled are matched

Then

$$\widehat{OR}_U = (\widehat{OR}_C)^2$$

↑ ↑
biased correct

e.g., $\widehat{OR}_C = 3 \Rightarrow \widehat{OR}_U = (3)^2 = 9$

R-to-1 matching



unconditional is overestimate of (correct) conditional estimate

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{OR}_U , is the square of the estimate obtained by using conditional ML estimation, which we denote by \widehat{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.

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 covariates 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

$L = L(\theta)$ = likelihood function

$$\theta = (\theta_1, \theta_2, \dots, \theta_q)$$

E, V, W model:

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

$$\theta = (\alpha, \beta, \gamma_1, \gamma_2, \dots, \delta_1, \delta_2, \dots)$$

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 $L(\theta)$, where θ denotes the collection of unknown parameters being estimated in the model. In matrix terminology, the collection θ is referred to as a **vector**; its components are the individual parameters being estimated in the model, denoted here as $\theta_1, \theta_2, \dots, \theta_q$, where q is the number of individual components.

For example, using the *E, V, W* logistic model previously described and shown here again, the unknown parameters are α, β , the γ_i 's, and the δ_j 's. Thus, the vector of parameters θ has α, β , the γ_i 's, and the δ_j 's as its components.

$L = L(\theta)$
= joint probability of observing the data

The likelihood function L or $L(\theta)$ 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

$\Pr(75 \text{ successes out of } 100 \text{ trials})$
has binomial distribution

$$\Pr(X = 75 | n = 100, p)$$

↑
given

$$\begin{aligned}\Pr(X = 75 | n = 100, p) \\ = c \times p^{75} \times (1 - p)^{100 - 75} \\ = L(p)\end{aligned}$$

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.

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.”

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 $L(\theta)$

$\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_q)$ = 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 θ which maximizes the likelihood function $L(\theta)$. The estimator is denoted as $\hat{\theta}$ and its components are $\hat{\theta}_1, \hat{\theta}_2$, and so on up through $\hat{\theta}_q$.

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 (Binomial)

ML solution:
 \hat{p} maximizes
 $L(p) = c \times p^{75} \times (1 - p)^{25}$

EXAMPLE (continued)

Maximum value obtained by solving

$$\frac{dL}{dp} = 0$$

for p :

$$\hat{p} = 0.75 \quad \text{"most likely"}$$

$$\begin{aligned} & \text{maximum} \\ & \downarrow \\ p > \hat{p} = 0.75 \Rightarrow L(p) < L(p = 0.75) \\ \text{e.g.,} & \quad \text{binomial formula} \\ p = 1 \Rightarrow L(1) &= c \times 1^{75} \times (1-1)^{25} \\ &= 0 < L(0.75) \end{aligned}$$

$$\hat{p} = 0.75 = \frac{75}{100}, \text{ a sample proportion}$$

Binomial model

$$\Rightarrow \hat{p} = \frac{X}{n} \text{ is ML estimator}$$

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 which 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 $L(\theta)$ is equivalent to maximizing $\ln L(\theta)$

$$\text{Solve: } \frac{\partial \ln L(\theta)}{\partial \theta_j} = 0, j = 1, 2, \dots, q$$

q equations in q unknowns require *iterative* solution by computer

In general, maximizing the likelihood function $L(\theta)$ is equivalent to maximizing the natural log of $L(\theta)$, which is computationally easier. The components of θ 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 j th individual parameter.

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 program (L_U)
vs.
conditional program (L_C)

likelihoods

Formula for L is built into computer programs

User inputs data and computer does calculations

L formulae are different for unconditional and conditional methods

The unconditional formula:
(a joint probability)

$$L_U = \prod_{l=1}^{m_1} P(\mathbf{X}_l) \prod_{l=m_1+1}^n [1 - P(\mathbf{X}_l)]$$

cases noncases
↓ ↓

$P(\mathbf{X})$ = logistic model

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

$$L_U = \frac{\prod_{l=1}^n \exp\left(\alpha + \sum_{i=1}^k \beta_i X_{il}\right)}{\prod_{l=1}^n \left[1 + \exp\left(\alpha + \sum_{i=1}^k \beta_i X_{il}\right)\right]}$$

As described earlier, if the model is logistic, there are **two alternative types** of computer programs to choose from, an **unconditional** versus a **conditional** program. These programs use different likelihood functions, namely, L_U for the unconditional method and L_C for the conditional method.

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 programs. 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 program does the heavy calculations of forming the likelihood function internally and maximizing this function to obtain the ML solutions.

Although we do not want to emphasize the particular likelihood formulae for the unconditional versus conditional methods, we do want to describe how these formulae are different. Thus, we briefly show these formulae for this purpose.

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 non-cases** (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 l th 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 l th noncase is given by $1 - P(\mathbf{X}_l)$.

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 :

$$L_C = \frac{\Pr(\text{observed data})}{\Pr(\text{all possible configurations})}$$

m_1 cases: $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{m_1})$

$n-m_1$ noncases: $(\mathbf{X}_{m_1+1}, \mathbf{X}_{m_1+2}, \dots, \mathbf{X}_n)$

$L_C = \Pr(\text{first } m_1 \text{ } \mathbf{X}'\text{s are cases} \mid \text{all possible configurations of } \mathbf{X}'\text{s})$

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 \mathbf{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 .

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.

The term **configuration** here refers to one of the possible ways that the observed set of \mathbf{X} vectors can be partitioned into m_1 cases and $n-m_1$ noncases. In example 1 here, for instance, the last m_1 \mathbf{X} vectors are the cases and the remaining \mathbf{X} 's are noncases. In example 2, however, the m_1 cases are in the middle of the listing of all \mathbf{X} vectors.

EXAMPLE: Configurations

- (1) Last m_1 \mathbf{X} 's are cases
 $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$
 _____ cases
- (2) Cases of \mathbf{X} 's are in middle of listing
 $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$
 _____ cases

Possible configurations

$$\begin{aligned} &= \text{combinations of } n \text{ things taken } m_1 \text{ at a time} \\ &= C_{m_1}^n \end{aligned}$$

$$L_C = \frac{\prod_{l=1}^{m_1} P(\mathbf{X}_l) \prod_{l=m_1+1}^n [1 - P(\mathbf{X}_l)]}{\sum_u \left\{ \prod_{l=1}^{m_1} P(\mathbf{X}_{ul}) \prod_{l=m_1+1}^n [1 - P(\mathbf{X}_{ul})] \right\}}$$

versus

$$L_U = \prod_{l=1}^{m_1} P(\mathbf{X}_l) \prod_{l=m_1+1}^n [1 - P(\mathbf{X}_l)]$$

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.

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 n observations into m_1 cases and $n-m_1$ noncases. Each configuration is indicated by the u in the L_C formula.

$$L_C = \frac{\prod_{l=1}^{m_1} \exp\left(\sum_{i=1}^k \beta_i X_{li}\right)}{\sum_u \left[\prod_{l=1}^{m_1} \exp\left(\sum_{i=1}^k \beta_i X_{lui}\right) \right]}$$

Note: α drops out of L_C

Conditional program:

- estimate β 's
- does not estimate α (nuisance parameter)

Note: OR involves only β 's

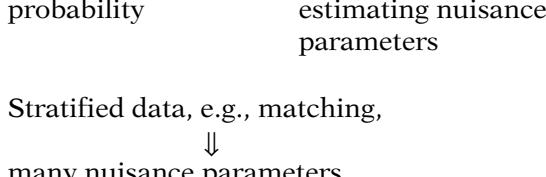
Case-control study: cannot estimate α

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.

The removal of the intercept α from the conditional likelihood is important because it means that when a conditional ML program 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 β 's and not α , we usually do not care about estimating α and, therefore, consider α to be a nuisance parameter.

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.

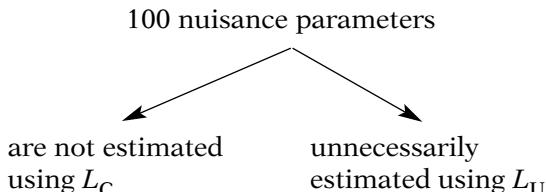
Regarding likelihood functions, then, we have shown that the unconditional and conditional likelihood functions involve different formulae. The unconditional formula has the theoretical advantage in that it is developed directly as a joint probability of the observed data. The conditional formula has the advantage that it does not require estimating nuisance parameters like α .



Stratified data, e.g., matching,



many nuisance parameters



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 \Rightarrow biased estimates of β 's
Conditional \Rightarrow unbiased estimates of β 's

If we consider the other parameters in the model for matched data, that is, the β 's, the unconditional likelihood approach gives biased estimates of the β 's, whereas the conditional approach gives unbiased estimates of the β 's.

V. Overview on Statistical Inferences for Logistic Regression

Chapter 5: Statistical Inferences Using Maximum Likelihood Techniques

Statistical inferences involve

- testing hypotheses
- obtaining confidence intervals

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.

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.

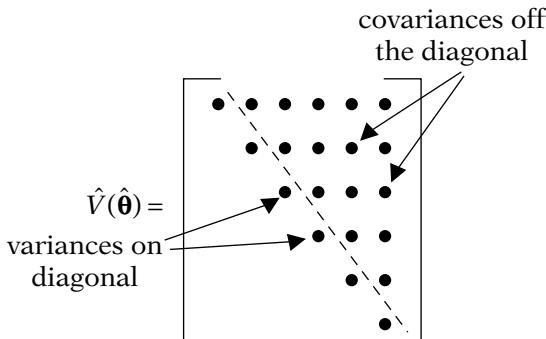
Quantities required from computer output:

- (1) Maximized likelihood value $L(\hat{\theta})$
- (2) Estimated variance-covariance matrix

Inference-making can be accomplished through the use of two quantities that are part of the output provided by standard ML estimation programs.

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 ($\hat{\theta}$) are substituted for their corresponding parameter values (θ). This value is called $L(\hat{\theta})$ in our earlier notation.

The second quantity is the **estimated variance-covariance matrix**. This matrix, \hat{V} of $\hat{\theta}$, 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.



Note: $\widehat{\text{cov}}(\hat{\theta}_1, \hat{\theta}_2) = r_{12}s_1s_2$

Importance of $\hat{V}(\hat{\theta})$:

inferences require accounting for variability and covariability

(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	$\hat{\beta}_k$	$s_{\hat{\beta}_k}$

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.

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 County, GA

$n = 609$ white males

9-year follow-up

$D = \text{CHD status}$

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
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

$$\text{CC} = \text{CAT} \times \text{CHL} \text{ and } \text{CH} = \text{CAT} \times \text{HPT}$$

→ $W's$ →

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 $\text{CC} = \text{CAT} \times \text{CHL}$ and $\text{CH} = \text{CAT} \times \text{HPT}$.

The exposure variable of interest here is the variable CAT, and the five covariates of interest, that is, the C's are AGE, CHL, ECG, SMK, and HPT. Using our E , V , W model framework described in the review section, we have E equals CAT, the five covariates equal to the V's, 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{OR} considers coefficients of CAT, CC, and CH

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

where

$$\begin{aligned}\hat{\beta} &= -12.6894 \\ \hat{\delta}_1 &= 0.0692 \\ \hat{\delta}_2 &= -2.3318\end{aligned}$$

$$\widehat{OR} = \exp[-12.6894 + 0.0692CHL + (-2.3318)HPT]$$

Must specify:

CHL and HPT
↑ ↓
effect modifiers

Note: \widehat{OR} different for different values specified for CHL and HPT

$$\widehat{OR} = \exp(-12.6894 + 0.0692CHL - 2.3318HPT)$$

		HPT	
		0	1
CHL	200	3.16	0.31
	220	12.61	1.22
	240	50.33	4.89

$$CHL = 200, HPT = 0: \quad \widehat{OR} = 3.16$$

$$CHL = 220, HPT = 1: \quad \widehat{OR} = 1.22$$

\widehat{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 \times CHL, and CH equal to CAT \times 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 covariates AGE, CHL, ECG, SMK, and HPT because each of the covariates is contained in the model as V variables.

$\widehat{\text{OR}}$'s = point estimators

Variability of $\widehat{\text{OR}}$ considered for statistical inferences

Two types of inferences:

- (1) testing hypotheses
- (2) interval estimation

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.

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.

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.

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 which makes use of maximized likelihood values such as 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.

EXAMPLES

- (1) Test for H_0 : $\text{OR} = 1$
- (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

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.

Large samples: both procedures give approximately the same results

Small or moderate samples: different results possible; likelihood ratio test preferred

Confidence intervals

- use large sample formulae
- use variance–covariance matrix

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 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.

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.

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.

EXAMPLE $\hat{V}(\hat{\theta})$					
	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

No interaction: variance only

Interaction: variances and covariances

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 104)

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 104–105)

- 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 versus conditional methods (pages 105–109)

- 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 109–115)

- A. $L = L(\theta)$ = likelihood function; gives joint probability of observing the data as a function of the set of unknown parameters given by $\theta = (\theta_1, \theta_2, \dots, \theta_q)$.
- B. ML method maximizes the likelihood function $L(\theta)$.
- 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 programs.
- 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 program estimates β 's 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 115–119)

- A. Two types of inferences: testing hypotheses and confidence interval estimation.
- B. Three items obtained from computer output for inferences:
 - i. Maximized likelihood value $L(\hat{\theta})$;
 - ii. Estimated variance–covariance matrix $\hat{V}(\hat{\theta})$: 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 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 $L(\theta)$ represents the joint probability of observing the data that has been collected for analysis.
- T F 7. The maximum likelihood method maximizes the function $\ln L(\theta)$.
- T F 8. The likelihood function formulae for both the unconditional and conditional approaches are the same.
- T F 9. The maximized likelihood value $L(\hat{\theta})$ 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 1200 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 10 parameters, then the ML solution solves a system of 10 equations in 10 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 lifetime sexual 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.	OR	P	95% Confidence Interval
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	
Introduction	126
Abbreviated Outline	126
Objectives	127
Presentation	128
Detailed Outline	150
Practice Exercises	152
Test	156
Answers To Practice Exercises	158

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 128)
- II. Information for making statistical inferences (pages 128–129)
- III. Models for inference-making (pages 129–130)
- IV. The likelihood ratio test (pages 130–134)
- V. The Wald test (pages 134–136)
- VI. Interval estimation: one coefficient (pages 136–138)
- VII. Interval estimation: interaction (pages 138–142)
- VIII. Numerical example (pages 142–149)

Objectives

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

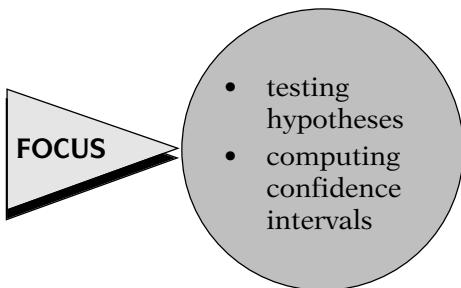
1. State the **null hypothesis** for testing the significance of a collection of one or more variables in terms of regression coefficients of a given logistic model.
2. Describe how to carry out a **likelihood ratio test** for the significance of one or more variables in a given logistic model.
3. Use 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

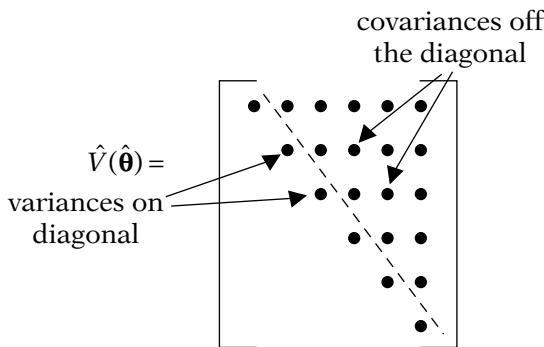
Quantities required from output:

- (1) Maximized likelihood value:

$$L(\hat{\theta})$$

- (2) Estimated variance-covariance matrix:

$$\hat{V}(\hat{\theta})$$



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.

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 $\hat{\theta}$ in our earlier notation.

The second quantity is the **estimated variance–covariance matrix**, which we denote as \hat{V} of $\hat{\theta}$.

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 pairs of ML estimates.

$$\widehat{\text{cov}}(\hat{\theta}_1, \hat{\theta}_2) = r_{12}s_1s_2$$

Importance of $\hat{V}(\hat{\theta})$:

Inferences require variances and covariances

(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	$\hat{\beta}_k$	$s_{\hat{\beta}_k}$

The reader may recall that the covariance between two estimates is the correlation times the standard errors of each estimate.

The variance-covariance matrix is important because hypothesis testing and confidence interval estimation require variances and sometimes covariances for computation.

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: $\text{logit } P_1(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2$

Model 2: $\text{logit } P_2(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Model 3: $\text{logit } P_3(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3 + \beta_5 X_2 X_3$

$\hat{L}_1, \hat{L}_2, \hat{L}_3$ are \hat{L} 's for models 1–3

$$\hat{L}_1 \leq \hat{L}_2 \leq \hat{L}_3$$

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 $X_1 X_3$ and $X_2 X_3$.

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.

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

$-2 \ln L_1 - (-2 \ln L_2) = LR$
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 chi-square 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.

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.

Model 1 special case of Model 2
Model 2 special case of Model 3

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

LR statistic (like F statistic) compares two models:

full model = larger model

reduced model = smaller model

H_0 : parameters in full model equal to zero

df = number of parameters set equal to zero

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.

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.

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.

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 .

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.

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

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)

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 confounders.

Then $\text{OR} = e^{\beta_3}$

$$H_0 : \beta_3 = 0 \Leftrightarrow H_0 : \text{OR} = e^0 = 1$$

$$\text{LR} = -2 \ln \hat{L}_1 - (-2 \ln \hat{L}_2)$$

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

If \hat{L}_2 much larger than \hat{L}_1 , then

$$\frac{\hat{L}_1}{\hat{L}_2} \approx 0$$

[Note : \ln_e (fraction) = negative]

$$\Rightarrow \ln \left(\frac{\hat{L}_1}{\hat{L}_2} \right) \approx \ln(0) = -\infty$$

$$\Rightarrow \text{LR} = -2 \ln \left(\frac{\hat{L}_1}{\hat{L}_2} \right) \approx \infty$$

Thus, X_3 highly significant \Rightarrow LR large and positive

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 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.

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 $-\infty$.

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

$$\hat{L}_2 \approx \hat{L}_1$$

$$\Rightarrow \frac{\hat{L}_1}{\hat{L}_2} \approx 1$$

$$\Rightarrow \text{LR} \approx -2 \ln(1) = -2 \times 0 = 0$$

Thus, X_3 nonsignificant $\Rightarrow \text{LR} \approx 0$

$$0 \leq \text{LR} \leq \infty$$

$$\uparrow \quad \uparrow$$

N.S. S.

Similar to chi square (χ^2)

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 .

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.

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.

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.

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

Model 2: $\text{logit } P_2(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
(reduced model)

Model 3: $\text{logit } P_3(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3 + \beta_5 X_2 X_3$
(full model)

EXAMPLE (continued)

$$H_0: \beta_4 = \beta_5 = 0$$

(similar to multiple-partial F test)

$$H_A: \beta_4 \text{ and/or } \beta_5 \text{ are not zero}$$

$$X_3 = E$$

X_1, X_2 confounders

X_1X_3, X_2X_3 interaction terms

$$H_0: \beta_4 = \beta_5 = 0 \Leftrightarrow H_0: \text{no interaction with } E$$

$$LR = -2 \ln \hat{L}_2 - (-2 \ln \hat{L}_3) = -2 \ln \left(\frac{\hat{L}_2}{\hat{L}_3} \right)$$

which is approx. χ^2 with two df under

$$H_0: \beta_4 = \beta_5 = 0$$

$$-2 \ln \hat{L}_2, -2 \ln \hat{L}_3$$



computer prints these
separately

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.

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 .

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.

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

$$\text{e.g., } H_0: \beta_3 = 0$$

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):

$$Z = \frac{\hat{\beta}}{s\hat{\beta}} \text{ is approximately } N(0, 1)$$

or

Z^2 is approximately χ^2 with 1 df

Variable	ML Coefficient	S.E.	Chi sq χ^2	P
X_1	$\hat{\beta}_1$	$s\hat{\beta}_1$	•	•
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
X_j	$\hat{\beta}_j$	$s\hat{\beta}_j$	•	•
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
X_k	$\hat{\beta}_k$	$s\hat{\beta}_k$	•	•

$LR \approx Z_{Wald}^2$ in large samples

$LR \neq Z_{Wald}^2$ in small to moderate samples

LR preferred (statistical)

Wald convenient—fit only one model

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.

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 chi-square 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.

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.

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.

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 to percentage points from a standard normal table.

EXAMPLE

Model 1: logit $P_1(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2$

Model 2: logit $P_2(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

$H_0 : \beta_3 = 0$

$$Z = \frac{\hat{\beta}_3}{s\hat{\beta}_3} \text{ is approx. } N(0, 1)$$

EXAMPLE (continued)

or

Z^2 is approximately χ^2 with one df

Wald test for more than one parameter:

requires matrices

(see *Epidemiologic Research*, Chapter 20, p. 431)

Or, alternatively, the Z can be squared and then compared to percentage points from a chi-square distribution with one degree of freedom.

Third testing method:

Score statistic

(see Kleinbaum et al., *Commun. in Statist.*, 1982)

The Wald test we have just described considers a null hypothesis involving only one model parameter. There is also a Wald test that considers null hypotheses involving more than one parameter, such as when comparing Models 2 and 3 above. However, 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 (*Epidemiologic Research*, Chapter 20, p. 431) for a description of this test.

Yet another method for testing these hypotheses involves the use of a **score statistic** (see Kleinbaum et al., *Communications in Statistics*, 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 chapter.

VI. Interval Estimation: One Coefficient

Large sample confidence interval:

estimate \pm (percentage point of Z \times 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% $\Rightarrow \alpha = 0.05$

$$\Rightarrow 1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975$$

$$Z_{0.975} = 1.96$$

CI for coefficient

vs.

✓ CI for odds ratio

As an example, if we focus on the β_3 parameter in Model 2, the 100 times $(1 - \alpha)\%$ confidence interval formula is given by $\hat{\beta}_3$ plus or minus the corresponding $(1 - \alpha/2)$ 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, $1 - \alpha/2$ is $1 - 0.025$ or 0.975, and $Z_{0.975}$ is equal to 1.96.

EXAMPLE

logit $P_2(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

$X_3 = (0, 1)$ variable

$$\Rightarrow \text{OR} = e^{\beta_3}$$

CI for OR: $\exp(\text{CI for } \beta_3)$

Model 2: $X_3 = (0, 1)$ exposure

X_1 and X_2 confounders

95% CI for OR:

$$\exp(\hat{\beta}_3 \pm 1.96 s_{\hat{\beta}_3})$$

Above formula assumes X_3 is coded as (0, 1)

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.

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 Chapter 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 coded as $\begin{cases} -1 & \text{unexposed} \\ 1 & \text{exposed} \end{cases}$

$$\text{OR} = \exp[1 - (-1)\beta_3] = e^{2\beta_3}$$

95% CI:

$$\exp(2\hat{\beta}_3 \pm 1.96 \times 2s_{\hat{\beta}_3})$$

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

$$\text{logit } P_3(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3 + \beta_5 X_2 X_3$$

$$\widehat{\text{OR}} = \exp(\hat{\beta}_3 + \hat{\beta}_4 X_1 + \hat{\beta}_5 X_2)$$

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 $X_1 X_3$ and $X_2 X_3$ in the model.

EXAMPLE

i.e., $\widehat{OR} = e^{\hat{l}}$,

where

$$l = \beta_3 + \beta_4 X_1 + \beta_5 X_2$$

100 $(1-\alpha)\%$ CI for e^l
similar to CI formula for e^{β_3}

$$\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{var}}(\hat{l})}\right]$$

$$\text{similar to } \exp\left[\hat{\beta}_3 \pm Z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{var}}(\hat{\beta}_3)}\right]$$

$\sqrt{\widehat{\text{var}}(\bullet)}$ = standard error

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.

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 .

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.

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.

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 denote the values of the i th 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 .

General CI formula :

$$\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{var}}(\hat{l})}\right]$$

$$\text{example : } l = \beta_3 + \beta_4 X_1 + \beta_5 X_2$$

General expression for l :

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

$OR = e^l$ where

$$l = \sum_{i=1}^k \beta_i (X_{1i} - X_{0i})$$

Interaction: variance calculation difficult

No interaction: variance directly from printout

$$\text{var}(\hat{l}) = \text{var}\left[\underbrace{\sum \hat{\beta}_i(X_{1i} - X_{0i})}_{\text{linear sum}}\right]$$

$\hat{\beta}_i$ are correlated for different i

Must use $\text{var}(\hat{\beta}_i)$ and $\text{cov}(\hat{\beta}_i, \hat{\beta}_j)$

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.

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.

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 .

It can be shown that the estimated variance of this linear function is given by the formula shown here.

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 (model 3)

$$\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{var}}(\hat{l})}\right],$$

$$\text{where } \hat{l} = \hat{\beta}_3 + \hat{\beta}_4 X_1 + \hat{\beta}_5 X_2$$

$$\begin{aligned} \widehat{\text{var}}(\hat{l}) &= \widehat{\text{var}}(\hat{\beta}_3) + (X_1)^2 \widehat{\text{var}}(\hat{\beta}_4) + (X_2)^2 \widehat{\text{var}}(\hat{\beta}_5) \\ &\quad + 2X_1 \widehat{\text{cov}}(\hat{\beta}_3, \hat{\beta}_4) + 2X_2 \widehat{\text{cov}}(\hat{\beta}_3, \hat{\beta}_5) \\ &\quad + 2X_1 X_2 \widehat{\text{cov}}(\hat{\beta}_4, \hat{\beta}_5) \end{aligned}$$

$\text{var}(\hat{\beta}_i)$ and $\text{cov}(\hat{\beta}_i, \hat{\beta}_j)$ obtained from printout BUT must specify 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

Recommendation: use “typical” or “representative” values of X_1 and X_2
 e.g., \bar{X}_1 and \bar{X}_2 in quintiles

Some computer packages compute

$$\widehat{\text{var}}(\hat{l})$$

General CI formula for E , V , W model:

$$\widehat{\text{OR}} = e^{\hat{l}},$$

where

$$\hat{l} = \beta + \sum_{j=1}^{p_2} \delta_j W_j$$

$$\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{var}}(\hat{l})}\right],$$

where

$$\begin{aligned} \widehat{\text{var}}(\hat{l}) &= \widehat{\text{var}}(\hat{\beta}) + \sum_{j=1}^{p_2} W_j^2 \widehat{\text{var}}(\hat{\delta}_j) \\ &\quad + 2 \sum_{j=1}^{p_2} W_j \widehat{\text{cov}}(\hat{\beta}, \hat{\delta}_j) + 2 \sum_j \sum_k W_j W_k \widehat{\text{cov}}(\hat{\delta}_j, \hat{\delta}_k) \end{aligned}$$

Obtain $\widehat{\text{var}}$'s and $\widehat{\text{cov}}$'s from printout *but* must specify W 's

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.

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 for logistic regression do compute the estimated variance of linear functions like \hat{l} as part of the program options. (See Computer Appendix.)

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 described in the review. Recall that the estimated odds ratio for this model can be written as e to \hat{l} , where \hat{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.

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 W 's in the model.

EXAMPLE

E, V, W model (Model 3):

$$X_3 = E,$$

$$X_1 = V_1 = W_1$$

$$X_2 = V_2 = W_2$$

$$\hat{l} = \hat{\beta}_3 + \hat{\beta}_4 X_1 + \hat{\beta}_5 X_2$$

$$= \hat{\beta} + \hat{\delta}_1 W_1 + \hat{\delta}_2 W_2$$

$$\beta = \beta_3,$$

$$p_2 = 2, W_1 = X_1, W_2 = X_2,$$

$$\delta_1 = \beta_4, \text{ and } \delta_2 = \beta_5$$

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

$n = 609$

EXAMPLE

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

$$E = \text{CAT}$$

$$C's = \text{AGE}, \text{CHL}, \text{ECG}, \text{SMK}, \text{HPT}$$

(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
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

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.

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

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
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
CH	-2.3318	0.7427	9.86	0.0017
CC	0.0692	0.3316	23.20	0.0000

interaction

CH = CAT \times HPT and CC = CAT \times CHL
unconditional ML estimation

n = 609, # parameters = 9

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

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

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 covariates AGE, CHL, ECG, SMK and HPT. Using the E , V , W formulation, this model contains five V variables, namely, the covariates, and no W variables.

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 \times HPT and CC equals the product CAT \times 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.

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.

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.

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.

test statistic info. available?

LR	no
Wald	yes

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

$$Z = \frac{0.5978}{0.3520} = 1.70$$

$$Z^2 = CHISQ = 2.88$$

$$P = 0.0896$$

(Assumes two-tailed test)

usual question: OR > 1? (one-tailed)

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.

$$\text{one-tailed } P = \frac{\text{two-tailed } P}{2}$$

$$= \frac{0.0894}{2} = 0.0447$$

$P < 0.05 \Rightarrow$ significant at 5% level

EXAMPLE (continued)

$H_0: \beta = 0$
equivalent to
 $H_0: \text{adjusted OR} = 1$

Variable	Coefficient	S.E.	Chi sq	P
Intercept				
CAT				
AGE				
CHL	0.0088	0.0033	(7.18)	0.0074
:				
HPT			↑	
				not of interest

95% CI for adjusted OR:
First, 95% CI for β :

$$\hat{\beta} \pm 1.96 \times s_{\hat{\beta}}$$

$$0.5978 \pm 1.96 \times 0.3520$$

$$\text{CI limits for } \beta: (-0.09, 1.29)$$

$$\exp(\text{CI limits for } \beta)$$

$$= (e^{-0.09}, e^{1.29})$$

$$= (0.91, 3.63)$$

CI contains 1,
so
do not reject H_0
at
5% level (*two-tailed*)

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 covariates. An equivalent way to state this null hypothesis is that the odds ratio for the effect of CAT on CHD adjusted for the five covariates is equal to the null value of 1.

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 covariates, 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.

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.

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.

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?

Model B vs. Model A

LR test for interaction:

$$H_0 : \delta_1 = \delta_2 = 0$$

where δ 's are coefficients of interaction terms CC and CH in model B

Full Model	Reduced Model
model B	model A
(interaction)	(no interaction)

$$\begin{aligned} LR &= -2 \ln \hat{L}_{\text{model A}} - (-2 \ln \hat{L}_{\text{model B}}) \\ &= 400.39 - 347.23 \\ &= 53.16 \end{aligned}$$

df = 2

significant at .01 level

\widehat{OR} for interaction model (B):

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

$\hat{\beta} = -12.6894$ for CAT

$\hat{\delta}_1 = 0.0692$ for CC

$\hat{\delta}_2 = -2.3318$ for CH

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.

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.

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.

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.

From the printout information given on pages 142–143, 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.

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 \times CHL and CH is CAT \times 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 $\hat{\beta}$ (-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)

$$\widehat{OR} = \exp[\beta + \delta_1 CHL + \delta_2 HPT] \\ = \exp[-12.6894 + 0.0692 CHL + (-2.3318) HPT]$$

Must specify
CHL and HPT
 \uparrow \uparrow
effect modifiers

adjusted \widehat{OR} :

		HPT	
		0	1
CHL	200	3.16	0.31
	220	12.61	1.22
	240	50.33	4.89

CHL = 200, HPT = 0 $\Rightarrow \widehat{OR} = 3.16$

CHL = 220, HPT = 1 $\Rightarrow \widehat{OR} = 1.22$

\widehat{OR} adjusts for AGE, CHL, ECG, SMK, and HPT (V variables)

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. 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.

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.

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 are contained in the model as V variables.

Confidence intervals:

$$\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\text{var}(\hat{l})}\right]$$

where

$$\hat{l} = \beta + \sum_{j=1}^{p_2} \delta_j W_j$$

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

$$\begin{aligned}\widehat{\text{var}}(\hat{l}) &= \widehat{\text{var}}(\hat{\beta}) + (W_1)^2 \widehat{\text{var}}(\hat{\delta}_1) + (W_2)^2 \widehat{\text{var}}(\hat{\delta}_2) \\ &\quad + 2W_1 \widehat{\text{Cov}}(\hat{\beta}, \hat{\delta}_1) + 2W_2 \widehat{\text{Cov}}(\hat{\beta}, \hat{\delta}_2) \\ &\quad + 2W_1 W_2 \widehat{\text{Cov}}(\hat{\delta}_1, \hat{\delta}_2)\end{aligned}$$

$W_1 = \text{CHL}$, $W_2 = \text{HPT}$

$\text{CHL} = 220$, $\text{HPT} = 1$:

$$\begin{aligned}\hat{l} &= \hat{\beta} + \hat{\delta}_1(220) + \hat{\delta}_2(1) \\ &= 0.1960\end{aligned}$$

$$\widehat{\text{var}}(\hat{l}) = 0.2279$$

$$\left[\begin{array}{c|c|c} \widehat{\text{var}}\hat{\beta} = 9.6389 & \widehat{\text{var}}\hat{\delta}_1 = 0.0002 & \widehat{\text{var}}\hat{\delta}_2 = 0.5516 \\ \widehat{\text{cov}}(\hat{\beta}, \hat{\delta}_1) = -0.0437 & \widehat{\text{cov}}(\hat{\delta}_1, \hat{\delta}_2) = -0.0049 & \widehat{\text{cov}}(\hat{\beta}, \hat{\delta}_2) = -0.0016 \end{array} \right]$$

95% CI for adjusted OR:

$$\exp[0.1960 \pm 1.96 \sqrt{0.2279}]$$

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. 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 \Rightarrow estimates have large variances

HPT = 1:

$\widehat{OR} = 0.31$, CI: (0.10, .91) below 1

$\widehat{OR} = 1.22$, CI: (0.48, 3.10) includes 1

$\widehat{OR} = 4.89$, CI: (1.62, 14.52) above 1

	\widehat{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.

Chapter 6: Modeling Strategy Guidelines

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.

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 covariates.

Detailed Outline

I. Overview (page 128)

Focus

- testing hypotheses
- computing confidence intervals

II. Information for making statistical inferences (pages 128–129)

- A. Maximized likelihood value: $L(\hat{\theta})$.
- B. Estimated variance–covariance matrix: $\hat{V}(\hat{\theta})$ 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 129–130)

- A. Model 1: $\text{logit } P(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2;$
 Model 2: $\text{logit } P(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3;$
 Model 3: $\text{logit } P(\mathbf{X}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3 + \beta_5 X_2 X_3.$
- 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 130–134)

- A. LR statistic compares two models: full (larger) model versus 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 versus Model 2: $\text{LR} = -2 \ln \hat{L}_1 - (-2 \ln \hat{L}_2)$, where $H_0: \beta_3 = 0$. This LR has approximately a chi-square distribution with one df under the null hypothesis.

$$\text{E. } -2 \ln \hat{L}_1 - (-2 \ln \hat{L}_2) = -2 \ln \left(\frac{\hat{L}_1}{\hat{L}_2} \right)$$

where $\frac{\hat{L}_1}{\hat{L}_2}$ 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 versus Model 3: $\text{LR} = -2 \ln \hat{L}_2 - (-2 \ln \hat{L}_3)$, where $H_0: \beta_4 = \beta_5 = 0$. This LR has approximately a chi-square distribution with two 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 134–136)

- A. Requires one parameter only to be tested, e.g., $H_0: \beta_3 = 0$.
- B. Test statistic: $Z = \frac{\hat{\beta}_3}{s_{\hat{\beta}_3}}$ 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: $Z = \frac{\hat{\beta}_3}{s_{\hat{\beta}_3}}$.

VI. Interval estimation: one coefficient (pages 136–138)

- A. Large sample confidence interval:
estimate \pm percentage point of $Z \times$ 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
$$\exp(\hat{\beta}_3 \pm 1.96 s_{\hat{\beta}_3})$$
- D. If X_3 has coding other than (0, 1), the CI formula must be modified.

VII. Interval estimation: interaction (pages 138–142)

- 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
$$\begin{aligned}\widehat{\text{var}}(\hat{l}) &= \widehat{\text{var}}(\hat{\beta}_3) + (X_1)^2 \widehat{\text{var}}(\hat{\beta}_4) + (X_2)^2 \widehat{\text{var}}(\hat{\beta}_5) \\ &\quad + 2X_1 \widehat{\text{cov}}(\hat{\beta}_3, \hat{\beta}_4) + 2X_2 \widehat{\text{cov}}(\hat{\beta}_3, \hat{\beta}_5) + 2X_1 X_2 \widehat{\text{cov}}(\hat{\beta}_4, \hat{\beta}_5).\end{aligned}$$

- B. General $100(1 - \alpha)\%$ CI formula for OR:

$$\exp\left[\hat{l} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{var}}(\hat{l})}\right]$$

where $\widehat{\text{OR}} = e^{\hat{l}}$,

$$\hat{l} = \sum_{i=1}^k \hat{\beta}_i (X_{1i} - X_{0i}) \text{ and } \text{var}(\hat{l}) = \text{var}\left(\underbrace{\sum \hat{\beta}_i (X_{1i} - X_{0i})}_{\text{linear sum}}\right)$$

- C. $100(1 - \alpha)\%$ 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 $\widehat{\text{OR}} = e^{\hat{l}}, \hat{l} = \hat{\beta} + \sum_{j=1}^{p_2} \hat{\delta}_j W_j$

$$\text{and } \widehat{\text{var}}(\hat{l}) = \widehat{\text{var}}(\hat{\beta}) + \sum_{j=1}^{p_2} W_j^2 \widehat{\text{var}}(\hat{\delta}_j) + 2 \sum_{j=1}^{p_2} W_j \widehat{\text{cov}}(\hat{\beta}, \hat{\delta}_j) + 2 \sum_j \sum_k W_j W_k \widehat{\text{cov}}(\hat{\delta}_j, \hat{\delta}_k)$$

- 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 142–149)

- 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 (McLawns 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 $SMK \times NS$ 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 $SMK \times NS$, 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. $H_0: \beta_6 = \beta_7 = \dots = \beta_{15} = 0$, 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 $H_0: \text{OR} = 1$, 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: $Z = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$, where $\hat{\beta}_1$ is the coefficient of HT in the no interaction model; $s_{\hat{\beta}_1}$

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:

$$\exp\left[\hat{\beta}_1 \pm 1.96\sqrt{\text{var}(\hat{\beta}_1)}\right],$$

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 $Z = \frac{\hat{\beta}_{\text{CAT}}}{s_{\hat{\beta}_{\text{CAT}}}}$, 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: $Z = \frac{0.5978}{0.3520} = 1.70$; 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

$$\begin{aligned}\exp\left[\hat{\beta}_{\text{CAT}} \pm 1.96\sqrt{\widehat{\text{var}}(\hat{\beta}_{\text{CAT}})}\right] &= (0.5978 \pm 1.96 \times 0.3520) \\ &= \exp(0.5978 \pm 0.6899) \\ &= (e^{-0.0921}, e^{1.2876}) = (0.91, 3.62).\end{aligned}$$

10. The null hypothesis for the likelihood ratio test for the effect of CC: $H_0: \beta_{CC} = 0$ where β_{CC} is the coefficient of CC in model II.
 Likelihood ratio statistic: $LR = -2 \ln \hat{L}_I - (-2 \ln \hat{L}_{II})$ where \hat{L}_I and \hat{L}_{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: $H_0: \beta_{CC} = 0$, where β_{CC} is the coefficient of CC in model II.
 Wald statistic: $Z = \frac{\hat{\beta}_{CC}}{s_{\hat{\beta}_{CC}}}$, 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: $Z = \frac{0.0683}{0.0143} = 4.77$; 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

$$\widehat{OR}_{adj} = \exp(\hat{\beta}_{CAT} + \hat{\delta}_{CC} \text{CHL}) = \exp(-14.0809 + 0.0683 \text{CHL}),$$

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

$$\text{CHL} = 220: \exp[-14.0809 + 0.0683(220)] = \exp(0.9451) = 2.57;$$

$$\text{CHL} = 240: \exp[-14.0809 + 0.0683(240)] = \exp(2.3111) = 10.09.$$

14. Formula for the 95% confidence interval for the adjusted odds ratio when CHL = 220:

$$\exp\left[\hat{l} \pm 1.96 \sqrt{\widehat{\text{var}}(\hat{l})}\right], \text{ where } \hat{l} = \hat{\beta}_{CAT} + \hat{\delta}_{CC}(220)$$

$$\text{and } \widehat{\text{var}}(\hat{l}) = \widehat{\text{var}}(\hat{\beta}_{CAT}) + (220)^2 \widehat{\text{var}}(\hat{\delta}_{CC}) + 2(220) \widehat{\text{cov}}(\hat{\beta}_{CAT}, \hat{\delta}_{CC})$$

where $\widehat{\text{var}}(\hat{\beta}_{CAT})$, $\widehat{\text{var}}(\hat{\delta}_{CC})$, and $\widehat{\text{cov}}(\hat{\beta}_{CAT}, \hat{\delta}_{CC})$ are obtained from the printout of the variance-covariance matrix.

6

Modeling

Strategy

Guidelines

Contents	
Introduction	162
Abbreviated Outline	162
Objectives	163
Presentation	164
Detailed Outline	183
Practice Exercises	184
Test	186
Answers to Practice Exercises	188

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 which 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. The initial model has to 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 (Chapter 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.

- I. Overview (page 164)
- II. Rationale for a modeling strategy (pages 164–165)
- III. Overview of recommended strategy (pages 165–169)
- IV. Variable specification stage (pages 169–171)
- V. Hierarchically well-formulated models (pages 171–173)
- VI. The hierarchical backward elimination approach (page 174)
- VII. The Hierarchy principle for retaining variables (pages 175–176)
- VIII. An example (pages 177–181)

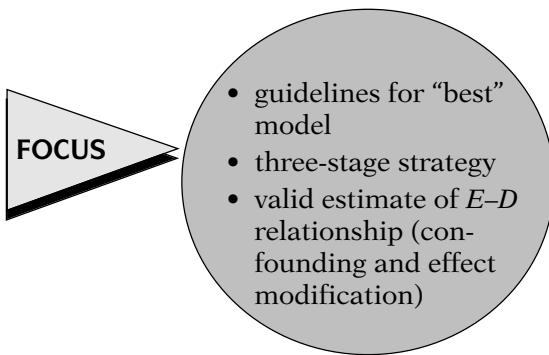
Objectives

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

1. State and recognize the three stages of the recommended modeling strategy.
2. Define and recognize a hierarchically well-formulated logistic model.
3. State, recognize, and apply the recommended strategy for choosing potential confounders in one's model.
4. State, recognize, and apply the recommended strategy for choosing potential effect modifiers in one's model.
5. State and recognize the rationale for a hierarchically well-formulated model.
6. State and apply the hierarchical backward elimination strategy.
7. State and apply the Hierarchy Principle.
8. State whether or not significance testing is allowed for the assessment of interaction and/or confounding.

Presentation

I. Overview



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

Minimum information in most study reports
e.g., little explanation about strategy

We begin by explaining the rationale for a modeling 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 chosen
- how variables selected
- how effect modifiers assessed
- how confounders assessed

Guidelines needed

- to assess validity of results
- to help researchers know what information to provide
- to encourage consistency in strategy
- for a variety of modeling procedures

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.

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

Two modeling goals
(1) to obtain a valid *E-D* estimate
(2) to obtain a good predictive model

(different strategies for different goals)

Prediction goal:
use computer algorithms

Validity goal

- our focus
- for etiologic research
- standard computer algorithms not appropriate

III. Overview of Recommended Strategy

Three stages
(1) variable specification
(2) interaction assessment
(3) confounding assessment followed by precision

Variable specification

- restricts attention to clinically or biologically meaningful variables
- provides largest possible initial model

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.

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.

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., (1998)).

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.

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 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

EXAMPLE

Suppose *gender* is effect modifier for *E-D* relationship:

$$\widehat{\text{OR}} \text{ males} = 5.4, \widehat{\text{OR}} \text{ females} = 1.2$$

interaction

Overall average = 3.5
not appropriate

Misleading because of separate effects
for males and females

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.

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

Interaction may not be of interest:

- skip interaction stage
- proceed directly to confounding

Thus, we recommend that if one wishes to assess interaction and also consider confounding, then the assessment of interaction comes first.

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

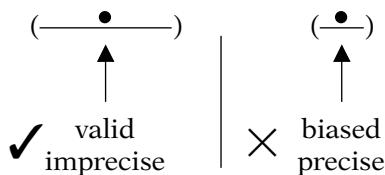
Study goal: single overall estimate. Then interaction not appropriate

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

Confounding followed by precision:



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.

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.

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.

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.

EXAMPLE

Control Variables	\hat{aOR}	95% CI
✓ AGE, RACE, SEX	2.4	(1.2, 3.7)
AGE	6.2	(5.9, 6.4)

VALID ↑ wide
BIASED ↑ narrow

$\frac{(1.2 \quad \bullet \quad 3.7)}{5.9 \quad \bullet \quad 6.4}$

VALIDITY BEFORE PRECISION
 ↓ ↓
 ✓ right answer precise answer

Confounding : *no statistical testing*



Validity—systematic error

(Statistical testing—random error)

Confounding in logistic regression—a validity issue

Computer algorithms no good (involve statistical testing)

Statistical issues beyond scope of this presentation:

- multicollinearity
- multiple testing
- influential observations

Multicollinearity

- independent variables approximately determined by other independent variables
- regression coefficients unreliable

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

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 which addresses systematic rather than random error. Statistical testing is appropriate for considering random error rather than systematic error.

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.

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 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 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

Specify D , E , C_1 , C_2 , ..., C_p based on

- study goals
- literature review
- theory

Specify V 's based on

- prior research or theory
- possible statistical problems

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.

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.

Next, we must specify the V 's, which are functions of the C 's that go into the model as potential confounders. Generally, we recommend that the choice of V 's be based primarily on prior research or theory, with some consideration of possible statistical problems like multicollinearity that might result from certain choices.

EXAMPLE

C 's: AGE, RACE, SEX

V 's:

Choice 1: AGE, RACE, SEX

Choice 2: AGE, RACE, SEX, AGE^2 ,
 $AGE \times RACE$, $RACE \times SEX$,
 $AGE \times SEX$

For example, if the C 's are AGE, RACE, and SEX, one choice for the V 's is the C 's themselves. Another choice includes AGE, RACE, and SEX plus more complicated functions such as AGE^2 , $AGE \times RACE$, $RACE \times SEX$, and $AGE \times SEX$.

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 if multicollinearity is found.

- ✓ Simplest choice for V 's
the C 's themselves (or a subset of C 's)

Specify W 's: (in model as $E \times W$)

restrict W 's to be V 's themselves or products of two V 's

(i.e., in model as $E \times V$ and $E \times V_i \times V_j$)

The simplest choice for the V 's is the C 's themselves. If the number of C 's is very large, it may even be appropriate to consider a smaller subset of the C 's considered to be most relevant and interpretable based on prior knowledge.

Once the V 's are chosen, the next step is to determine the W 's. 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 W 's be restricted either to the V 's themselves or to product terms involving two V 's. 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 V 's.

Most situations:

specify V 's and W 's as C 's or subset of C 's

EXAMPLE

$C_1, C_2, C_3, = AGE, RACE, SEX$

$V_1, V_2, V_3, = AGE, RACE, SEX$

W 's = subset of AGE, RACE, SEX

Rationale for W 's (common sense):

Product terms more complicated than EV_iV_j are

- difficult to interpret
- typically cause multicollinearity
- ✓ Simplest choice: use EV_i terms only

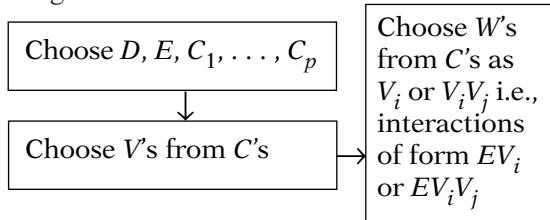
For most situations, we recommend that both the V 's and the W 's be the C 's themselves, or even a subset of the C 's.

As an example, if the C 's are AGE, RACE, and SEX, then a simple choice would have the V 's be AGE, RACE, and SEX and the W 's be a subset of AGE, RACE, and SEX thought to be biologically meaningful as effect modifiers.

The **rationale** for our recommendation about the W 's 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 **multicollinearity** 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



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 C 's, then the choice of the V 's from the C 's and, finally, the choice of the W 's in terms of the C 's.

V. Hierarchically Well-Formulated Models

Initial model structure: HWF

Model contains **all lower-order components**

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.

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.

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.

In contrast, the model given by 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 $\delta_2 EV_2$ 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

Not HWF model:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 + \delta_1 EV_1 + \delta_2 EV_2 + \delta_3 EV_1V_2$$

Components of EV_1V_2 :

$$E, V_1, V_2, EV_1, EV_2, V_1V_2$$

↑ not in model

EXAMPLE

HWF model:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 + \delta_1 EV_1 + \delta_2 EV_2$$

Components of EV_1 :

$$E, V_1 \text{ both in model}$$

EXAMPLE

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1^2 + \gamma_2 V_2 + \delta_1 EV_1^2$$

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 :

model does not contain

- V_1 , component of V_1^2
- EV_1 , component of EV_1^2

Why require HWF model?

Answer:

HWF? Tests for highest-order variables?

No	dependent on coding
Yes	independent of coding

For illustrative purposes, let us consider one other model given by logit $P(\mathbf{X})$ 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 .

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.

To illustrate this point, we return to the first example considered above, where the model is given by 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 $\delta_3 EV_1 V_2$. 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 $EV_1 V_2$.

EXAMPLE

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 + \delta_1 EV_1 + \delta_2 EV_2 + \delta_3 EV_1 V_2$$

Not HWF model:

$V_1 V_2$ missing

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 non-significant 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.

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.

EXAMPLE

HWF model:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 + \delta_3 V_1 V_2 \\ + \delta_1 EV_1 + \delta_2 EV_2 + \delta_3 EV_1 V_2$$

Testing for EV_1V_2 is **independent of coding** of E : $(0, 1)$, $(-1, 1)$, or other.

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.

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.

EXAMPLE

HWF model:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 + \gamma_3 V_1 V_2 \\ + \delta_1 EV_1 + \delta_2 EV_2 + \delta_3 EV_1 V_2$$

EV_1V_2 : **not dependent** on coding

EV_1 or EV_2 : **dependent** on 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.

VI. The Hierarchical Backward Elimination Approach

✓ Variable specification

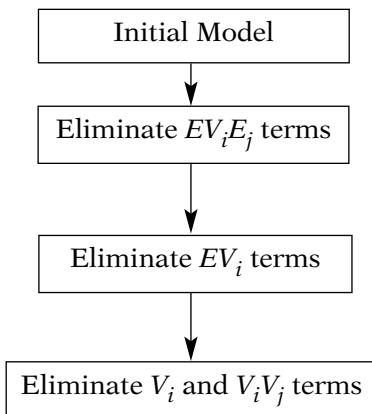
✓ HWF model

Largest model considered

= initial (starting) model

Initial model → Final model

hierarchical backward elimination



EV_i and EV_j (interactions):
use statistical testing

V_i and V_j (confounders):
do not use statistical testing

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 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 $EV_i V_j$, 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_j terms.

As described in later sections, the $EV_i V_j$ and EV_i product terms can be eliminated using appropriate statistical testing methods.

However, decisions about the V_i and V_j terms, which are potential confounders, should not involve statistical testing.

Hierarchical

3 factors: $EV_i V_j$
 \downarrow
 2 factors: EV_i
 \downarrow
 2 factors: $V_i V_j$
 \downarrow
 1 factor: V_i

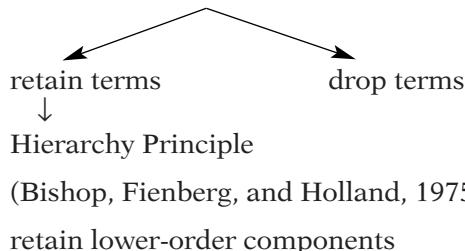
Backward

Large starting model
 \downarrow
 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.

VII. The Hierarchy Principle for Retaining Variables

Hierarchical Backward Elimination



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.

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_iV_j terms

Suppose: EV_2V_5 significant

Hierarchy 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 which 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.

Hierarchy Principle rationale

- tests for lower-order components depend on coding
- tests should be independent of coding
- therefore, no tests allowed 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 $(-1, 1)$

HWF model:

tests for *highest-order* terms *independent* of coding

but

tests for *lower-order* terms *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_i **dependent** on coding

EXAMPLE

HWF: EV_i highest-order terms

Then tests for

EV_i **independent** of coding **but** tests for V_i **dependent** on coding

Hierarchy Principle

- ensures that the model is HWF

e.g., EV_iV_j is significant \Rightarrow retain lower-order components or else model is not HWF

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.

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 $(-1, 1)$.

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.

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.

If the highest-order terms of a hierarchically well-formulated 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.

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.

VIII. An Example

EXAMPLE

Cardiovascular Disease Study
9-year follow-up Evans County, GA
 $n = 609$ white males

The variables:

CAT, AGE, CHL, SMK, ECG, HPT

 at start
CHD = outcome

CAT: (0, 1) exposure
AGE, CHL: continuous
SMK, ECG, HPT: (0, 1)

} control
variables

$E = \text{CAT} \rightarrow D = \text{CHD}$

controlling for

AGE, CHL, SMK, ECG, HPT

 C's

Variable specification stage:

V 's: potential confounders in initial model

Here, V 's = C 's:

$$\begin{aligned} V_1 &= \text{AGE}, V_2 = \text{CHL}, V_3 = \text{SMK}, \\ V_4 &= \text{ECG}, V_5 = \text{HPT} \end{aligned}$$

Other possible V 's:

$$V_6 = \text{AGE} \times \text{CHL}$$

$$V_7 = \text{AGE} \times \text{SMK}$$

$$V_8 = \text{AGE}^2$$

$$V_9 = \text{CHL}^2$$

We 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 6 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).

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.

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 C 's that we have specified at the start of our modeling strategy.

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.

In choosing the V 's, we follow our earlier recommendation to let the V 's be the same as the C 's. Thus, we will let $V_1 = \text{AGE}$, $V_2 = \text{CHL}$, $V_3 = \text{SMK}$, $V_4 = \text{ECG}$, and $V_5 = \text{HPT}$.

We could have chosen other V 's in addition to the five C 's. For example, we could have considered V 's which are products of two C 's, such as V_6 equals $\text{AGE} \times \text{CHL}$ or V_7 equals $\text{AGE} \times \text{SMK}$. We could also have considered V 's which are squared C 's, such as V_8 equals AGE^2 or V_9 equals CHL^2 .

EXAMPLE (continued)

Restriction of V 's to C 's because

- large number of C 's
- additional V 's difficult to interpret
- additional V 's may lead to collinearity

Choice of W 's:

(go into model as EW)

W 's = C 's:

$$\begin{aligned} W_1 &= \text{AGE}, W_2 = \text{CHL}, W_3 = \text{SMK}, \\ W_4 &= \text{ECG}, W_5 = \text{HPT} \end{aligned}$$

Other possible W 's:

$$W_6 = \text{AGE} \times \text{CHL}$$

(If W_6 is in model, then

$V_6 = \text{AGE} \times \text{CHL}$ also in HWF model.)

Alternative choice of W 's:

Subset of C 's, e.g.,

$$\text{AGE} \Rightarrow \text{CAT} \times \text{AGE} \text{ in model}$$

$$\text{ECG} \Rightarrow \text{CAT} \times \text{ECG} \text{ in model}$$

Rationale for W 's = C 's:

- 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_i 's = C 's = W_j 's

However, we have restricted the V 's to the C 's themselves primarily because there are a moderately large number of C 's being considered, and any further addition of V 's is likely to make the model difficult to interpret as well as difficult to fit because of likely collinearity problems.

We next choose the W 's, which are the variables that go into the initial model as product terms with $E(\text{CAT})$. These W 's are the potential effect modifiers to be considered. The W 's that we choose are the C 's themselves, which are also the V 's. That is, W_1 through W_5 equals AGE, CHL, SMK, ECG, and HPT, respectively.

We could have considered other choices for the W 's. For instance, we could have added two-way products of the form W_6 equals AGE \times 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 \times CHL, to make our model hierarchically well formulated. This is because AGE \times CHL is a lower-order component of CAT \times AGE \times CHL, which is EW_6 .

We could also have considered for our set of W 's some subset of the five C 's, rather than all five C 's. For instance, we might have chosen the W 's to be AGE and ECG, so that the corresponding product terms in the model are CAT \times AGE and CAT \times ECG only.

Nevertheless, we have chosen the W 's to be all five C 's so as to consider the possibility of interaction from any of the five C 's, 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?

e.g., $CAT \times AGE$

↓

CAT and AGE both in model as main effects

HWF model? **YES**

If $CAT \times ECG \times SMK$ in model, then

not HWF model

because

$ECG \times SMK$ not in model

Next

Hierarchical Backward Elimination Procedure

First, eliminate EW terms

Then, eliminate V terms

Interaction assessment
and
confounding assessments (details in
Chapter 7)

Results of Interaction Stage

$CAT \times CHL$ and $CAT \times HPT$
are the only two interaction terms to
remain in the model

Model contains

$CAT, AGE, CHL, SMK, ECG, HPT,$

$\underbrace{\hspace{10em}}$
 $V's$

$CAT \times CHL$ and $CAT \times HPT$

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.

For example, the lower-order components of a product variable like $CAT \times AGE$ 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.

Note that if we add to the above model the three-way product term $CAT \times ECG \times SMK$, the resulting model is not hierarchically well formulated. This is because the term $ECG \times SMK$ has not been specified as one of the V variables in the model.

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.

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.

Here, we summarize the results of the interaction assessment and confounding assessment stages and then return to provide more details of this example in Chapter 7.

The results of the interaction stage allow us to eliminate three interaction terms, leaving in the model the two product terms $CAT \times CHL$ and $CAT \times HPT$.

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 $CAT \times CHL$ and $CAT \times HPT$.

EXAMPLE (continued)

All five V 's in model so far

Hierarchy Principle

identify V 's that ***cannot*** be eliminated
 EV_i significant
 \downarrow
 E and V_i must remain

$CAT \times CHL \Rightarrow CAT$ and CHL remain
 $CAT \times HPT \Rightarrow CAT$ and HPT remain

Thus,

CAT (exposure) remains
plus
 CHL and HPT remain

AGE , SMK , ECG
eligible for elimination

Results (details in Chapter 7):

Cannot remove AGE , SMK , ECG
(decisions too subjective)

Final model variables:

CAT , AGE , CHL , SMK , ECG , HPT ,
 $CAT \times CHL$, and $CAT \times HPT$

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 $CAT \times CHL$ are CAT and CHL , and the lower-order components of $CAT \times HPT$ 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 Chapter 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 $CAT \times CHL$ and $CAT \times HPT$.

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
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

interaction

$$\begin{aligned} CH &= \text{CAT} \times \text{HPT} \text{ and} \\ CC &= \text{CAT} \times \text{CHL} \end{aligned}$$

$$\widehat{\text{ROR}} = \exp(-12.6894 + 0.0692\text{CHL} - 2.3881\text{HPT})$$

Details in Chapter 7.

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 \times HPT and CAT \times CHL are denoted in the printout as CH and CC, respectively.

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 Chapter 7.

SUMMARY

Three stages

- (1) variable specification
- (2) interaction
- (3) confounding/precision

Initial model: HWF model

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**.

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.

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.

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.

Hierarchy Principle

significant product term



retain lower-order components

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 164)

Focus:

- guidelines for “best” model
- 3-stage strategy
- valid estimate of E - D relationship

II. Rationale for a modeling strategy (pages 164–165)

- 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 165–169)

- 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 169–171)

- A. Start with D , E , and C_1, C_2, \dots, 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. Hierarchically well-formulated (HWF) models (pages 171–173)

- 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 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.

VI. The hierarchical backward elimination approach (page 174)

- A. Flow diagram representation.
- B. Flow description: evaluate $EV_i V_j$ 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.

VII. The Hierarchy Principle for retaining variables (pages 175–176)

- 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.
- B. Example.
- 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.
- D. Example.

VIII. An example (pages 177–181)

- A. Evans County CHD data description.
- B. Variable specification stage.
- C. Final results.

Practice Exercises

A prevalence study of predictors of surgical wound infection in 265 hospitals throughout Australia collected data on 12,742 surgical patients (McLawns 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 $HT \times AGE \times SEX$ 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 $HT \times AGE \times SEX$. 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 $HT \times AGE$. 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: $HT \times CT$

and $HT \times SEX$; 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 = 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

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:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta \text{SMK} + \sum \gamma_i^* V_i^* + \gamma_1 \text{NS} + \gamma_2 \text{AS} + \gamma_3 \text{NS} \times \text{AS} \\ + \delta_1 \text{SMK} \times \text{NS} + \delta_2 \text{SMK} \times \text{AS} + \delta_3 \text{SMK} \times \text{NS} \times \text{AS},$$

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 1, is a test for the significance of the three-factor product term $\text{SMK} \times \text{NS} \times \text{AS}$ dependent on the coding of SMK ? If so, explain why; if not explain, why not.
13. For the model in Question 1, is a test for the significance of the two-factor product term $\text{SMK} \times \text{NS}$ dependent on the coding of SMK ? If so, explain why; if not, explain why not.
14. For the model in Question 1, briefly describe a hierarchical backward elimination procedure for obtaining a best model.
15. Suppose that the three-factor product term $\text{SMK} \times \text{NS} \times \text{AS}$ 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., $\text{SMK} \times \text{NS} \times \text{AS}$ 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 $\text{HT} \times \text{AGE} \times \text{SEX}$ is added to the model, the new model will *not* be hierarchically well formulated because the lower-order component $\text{AGE} \times \text{SEX}$ is not contained in the original nor new model.
5. A test for $\text{HT} \times \text{AGE} \times \text{SEX}$ in the above model is dependent on coding in the sense that different test results (e.g., rejection versus 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 $\text{HT} \times \text{AGE}$ in the model of Exercise 3 is independent of coding because the model is hierarchically well formulated and the $\text{HT} \times \text{AGE}$ 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, $\text{HT} \times \text{AGE}$, 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 which 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 $HT \times CT$ and $HT \times SEX$ 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 appropriate to test for their inclusion in the model.
10. At the end of the interaction assessment stage, the remaining model is given by

$$\text{logit } P(\mathbf{X}) = \alpha + \beta HT + \gamma_1 HS + \gamma_2 CT + \gamma_3 AGE + \gamma_4 SEX + \delta_2 HT \times CT + \delta_4 HT \times SEX.$$

7

Modeling Strategy for Assessing Interaction and Confounding

Contents	
Introduction	192
Abbreviated Outline	192
Objectives	193
Presentation	194
Detailed Outline	221
Practice Exercises	222
Test	224
Answers to Practice Exercises	225

Introduction

This chapter continues the previous chapter (Chapter 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 (Chapter 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 which 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.

- I. Overview (pages 194–195)
- II. Interaction assessment stage (pages 195–198)
- III. Confounding and precision assessment when no interaction (pages 199–203)
- IV. Confounding assessment with interaction (pages 203–211)
- V. The Evans County example continued (pages 211–218)

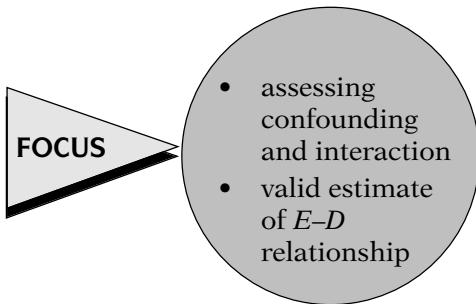
Objectives

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;
 - b. when there is interaction.

Presentation

I. Overview



Three stages

- (1) variable specification
- (2) interaction
- (3) confounding/precision

Initial model: HWF

$$\begin{array}{ll} EV_i V_j & EV_i, EV_j \\ \text{in initial} \rightarrow & V_i, V_j, V_i V_j \\ \text{model} & \text{also in model} \end{array}$$

Hierarchical backward elimination:

- can test for interaction, *but* not confounding
- can eliminate lower-order term if corresponding higher-order term is not significant

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.

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**.

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.

Thus, for example, if the model contains an interaction term of the form $EV_i V_j$, this will require the lower-order terms EV_i , EV_j , V_i , V_j , and $V_i V_j$ also to be in the initial model.

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.

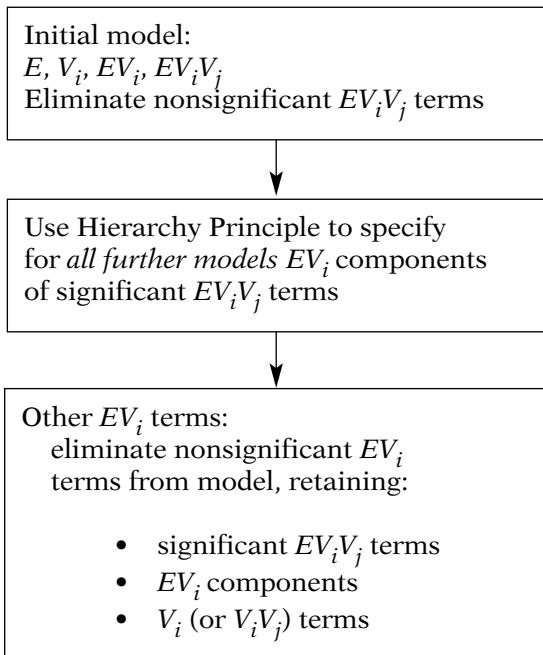
II. Interaction Assessment Stage

Start with HWF model

Use hierarchical backward elimination:

EV_iV_j before EV_i

Interaction stage flow:



Statistical testing

Chunk test for entire collection of interaction terms

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 .

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 which 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.

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

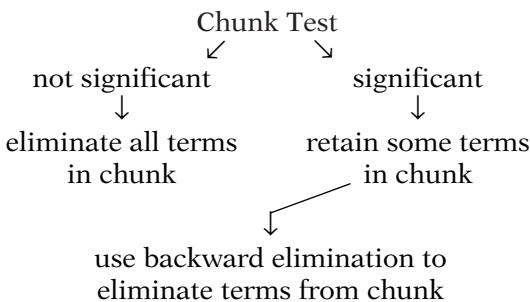
$EV_1V_2, EV_1V_3, EV_2V_3$ in model

chunk test for $H_0 : \delta_1 = \delta_2 = \delta_3 = 0$

use LR statistic $\sim \chi^2_3$ comparing

full model: all $V_i, V_iV_j, EV_i, EV_iV_j$ with
reduced model: V_i, V_iV_j, EV_j

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.



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 algorithm to eliminate insignificant variables from the model one at a time. Depending on the preference of the investigator, such a backward elimination procedure may be carried out without even doing the chunk test or regardless of the results of the chunk test.

EXAMPLE

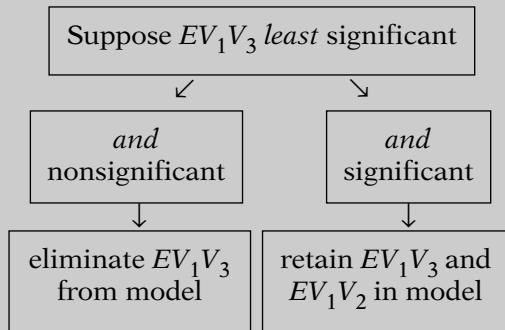
HWF model:

$$\begin{array}{c} \text{EV}_1\text{V}_2, \text{EV}_1\text{V}_3, \\ \hline \text{V}_1, \text{V}_2, \text{V}_3, \text{V}_1\text{V}_2, \text{V}_1\text{V}_3, \\ \text{EV}_1, \text{EV}_2, \text{EV}_3 \end{array}$$

As an example of such a backward algorithm, suppose we again consider a hierarchically well-formulated model which 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)

Backward approach:



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

Suppose EV_1V_2 significant:
then EV_1V_2 retained and above
reduced model is current model

Next: eliminate EV terms

From Hierarchy Principle:

E, V_1, V_2, EV_1, EV_2 , and V_1V_2
retained in all further models

Assess other EV_i terms:
 only EV_3 eligible for removal

Using the backward 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 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 is then fitted. The EV_1V_2 term is then dropped if nonsignificant but is retained if significant.

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.

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.

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: $EV_1V_2, EV_1, EV_2, (EV_3,$
 $V_1, V_2, V_3, V_1V_2, V_1V_3)$

Reduced model: $EV_1V_2, EV_1, EV_2,$
 $V_1, V_2, V_3, V_1V_2, V_1V_3$

$$\text{Wald test: } Z = \frac{\hat{\delta}_{EV_3}}{S_{\hat{\delta}_{EV_3}}}$$

Suppose both LR and Wald tests are nonsignificant:

then drop EV_3 from model

Interaction stage results:

$$\begin{array}{l} EV_1V_2, EV_1, EV_2 \\ V_1, V_2, V_3 \\ V_1V_2, V_1V_3 \end{array} \left. \begin{array}{l} \} \\ \} \end{array} \right. \text{confounders}$$

All V_i (and V_iV_j) remain in model after interaction assessment

Most situations
use *only* EV_i
product terms
 \Downarrow
interaction assessment
less complicated
 \Downarrow
do not need V_iV_j terms
for HWF model.

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 which 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 that both the above likelihood ratio and Wald tests are nonsignificant. Then we can drop the variable EV_3 from the model.

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 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 confounding terms in this model because they do not involve the exposure variable E .

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)

Confounding before Precision
 ↓ ↓
 ✓ gives correct answer gives narrow confidence interval

No interaction model:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \sum \gamma_i V_i$$

(no terms of form EW)

Interaction present?	Confounding assessment?
No	Straightforward
Yes	Difficult

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 which characterize statistical testing.

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.

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 .

The assessment of confounding is relatively straightforward 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.

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 which represents an adjusted estimate which controls for all five V variables.

EXAMPLE

Initial model

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \dots + \gamma_5 V_5$$

$$\widehat{\text{OR}} = e^{\hat{\beta}}$$

(a single number)
adjusts for V_1, \dots, V_5

EXAMPLE (continued)

Gold standard estimate:
controls for all potential confounders
(i.e., all five V 's)

Other OR estimates:
drop some V 's
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
correct answer

$$\widehat{\text{OR}} \text{ (reduced)} \stackrel{?}{=} \widehat{\text{OR}} \text{ (gold standard)}$$

If different, then reduced model *does not* control for confounding

Suppose:
 Gold standard (all five V 's)
 $\widehat{\text{OR}} = 2.5$
 reduced model (V_1 and V_2)
 \uparrow $\widehat{\text{OR}} = 5.2$
 does not control meaningfully
for confounding different

$$\widehat{\text{OR}} \left(\begin{array}{l} \text{some other} \\ \text{subset of } V\text{'s} \end{array} \right) \stackrel{?}{=} \widehat{\text{OR}} \left(\begin{array}{l} \text{gold} \\ \text{standard} \end{array} \right)$$

If equal, then subset controls confounding

$$\begin{aligned} \widehat{\text{OR}}(V_3 \text{ alone}) &= 2.7 \\ \widehat{\text{OR}}(V_4 \text{ and } V_5) &= 2.3 \\ \widehat{\text{OR}} \text{ (gold standard)} &= 2.5 \end{aligned}$$

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 V 's, in our model.

We can nevertheless obtain other estimated odds ratios by dropping some of the V 's 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 V 's.

Because the reduced model is different from the gold standard model, the estimated odds ratio obtained for the reduced model may be meaningful 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 V 's 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 V 's 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

$$\widehat{OR} = e^{\hat{\beta}}$$
- identify subsets of V 's giving approximately same \widehat{OR} as gold standard

If \widehat{OR} (subset of V 's) = \widehat{OR} (gold standard), then

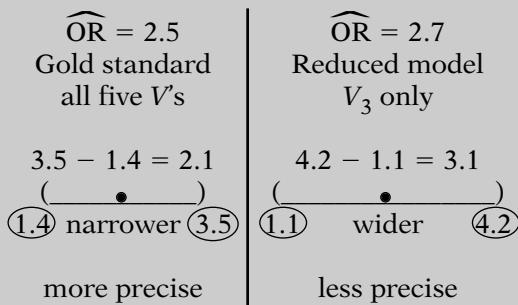
- which subset to use?
- why not use gold standard?

Answer: precision

less precise	more precise
CI's: (<u> • </u>)	(<u> • </u>)
less narrow	more narrow

EXAMPLE

95% confidence interval (CI)



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 we find one or more subset 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 don't we just use the gold standard?

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.

For example, suppose the 95% confidence interval around the gold standard \widehat{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 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.

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 don't we use gold standard?

Answer: Might find subset of V 's which will

- gain precision (narrower CI)
- without sacrificing validity (same point estimate)

The answer to the question, Why don't 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.

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 which uses V_4 and V_5 in our model to the gold standard estimate.

EXAMPLE

Model	\widehat{OR}	CI
$\checkmark V_4$ and V_5	same (2.3)	narrower (1.9, 3.1)
Gold standard	same (2.5)	wider (1.4, 3.5)

Which subset to control?

Answer: subset with most meaningful gain in precision

Eligible subset: same point estimate as 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

Scientific: Gold standard uses *all* relevant variables for control

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.

By **eligible subset**, we mean any collection of V 's that gives essentially the same point estimate as the 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.

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

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \cdots + \gamma_5 V_5$$

<u>V's in model</u>	$e^{\hat{\beta}}$	<u>95% CI</u>
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	*	

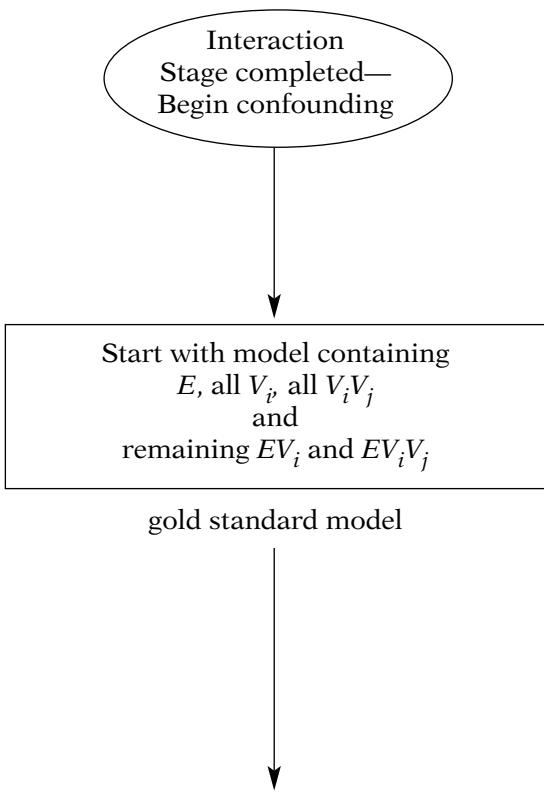


* $e^{\hat{\beta}}$ meaningfully different from 2.5

Returning to our example involving five V variables, suppose that the point estimates and confidence intervals for various subsets of V 's 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 V 's, that is, the gold standard, because no meaningful gain in precision is obtained from controlling for either of the two eligible subsets of V 's. 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.

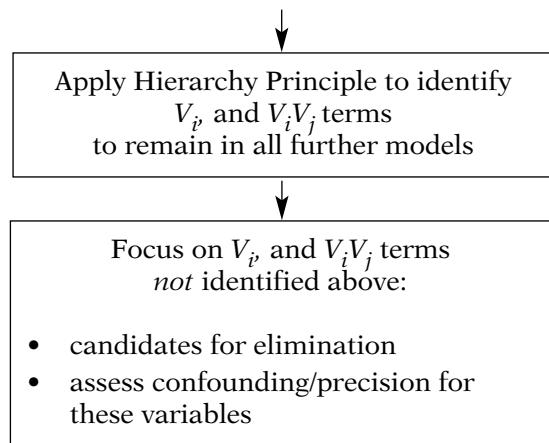
IV. Confounding Assessment with Interaction



We now consider how to assess confounding when the model contains interaction terms. A flow diagram which 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_i 's and V_iV_j 's.



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



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{OR} change?

gold standard vs. model without one
model or more V_i and V_iV_j

- (1) Identify subsets so that $\widehat{OR}_{\text{gold standard}} \approx \widehat{OR}_{\text{subset}}$
 - (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_i 's and V_iV_j 's.

More specifically, we carry out the following two steps:

- (1) Identify those subsets of V_i 's and V_iV_j 's giving approximately the same odds ratio estimate as the gold standard.
- (2) Control for that subset which gives the largest gain in precision.

$$\text{Interaction: } \widehat{\text{OR}} = \exp(\hat{\beta} + \sum \hat{\delta}_j W_j)$$

$\hat{\beta}$ and $\hat{\delta}_j$ nonzero

$$\text{no interaction: } \widehat{\text{OR}} = \exp(\hat{\beta})$$

Coefficients change when potential confounders dropped:

- meaningful change?
- subjective?

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 $\hat{\beta}$.

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:

$$\begin{aligned} E, V_1, V_2, V_3, V_4 &= V_1 V_2 \\ EV_1, EV_2, EV_3, EV_4 &= EV_1 V_2 \end{aligned}$$

Suppose EV_4 ($= EV_1 V_2$) significant

Hierarchy Principle:

EV_1 and EV_2 retained in all further models

EV_3 candidate to be dropped

Test for EV_3 (LR or Wald test)

V_1, V_2, V_3, V_4 (**all** potential confounders) forced into model during interaction stage

As an example, suppose our initial model contains E , four V 's, namely, V_1, V_2, V_3 , and $V_4 = V_1 V_2$, and four EV 's, 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 $EV_1 V_2$.

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.

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 $EV_1 V_2$. This leaves EV_3 as the only remaining two-factor interaction candidate to be dropped if not significant.

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 $E, V_1, V_2, V_3, V_4 = V_1 V_2$, EV_1, EV_2 and EV_4 are forced into the model.

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

$$E, \underbrace{V_1, V_2, V_3, V_4}_{V's}, \underbrace{EV_1, EV_2, EV_3, EV_4}_{EV's}$$

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 $H_0 : \beta_{EV_3} = 0$ **in full model**

Suppose EV_3 not significant



model after interaction assessment:

$$E, \underbrace{V_1, V_2, V_3, V_4}_{\text{where } V_4 = V_1 V_2}, EV_1, EV_2, EV_4$$

potential confounders

Hierarchy Principle:

identify V 's not eligible to be dropped—lower-order components

$EV_1 V_2$ significant

↓ Hierarchy Principle

Retain V_1, V_2 , and $V_4 = V_1 V_2$

Only V_3 eligible to be dropped

$$\widehat{\text{OR}}_{V_1, V_2, V_3, V_4} \stackrel{?}{=} \widehat{\text{OR}}_{V_1, V_2, V_4}$$

↑
excludes V_3

The likelihood ratio test for the significance of EV_3 compares a “full” model containing E , the four V 's, EV_1 , EV_2 , EV_3 , and EV_4 with a reduced model which 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 V 's, 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 $EV_1 V_2$ has been found significant, we must retain in all further models the lower-order components V_1 , V_2 , and $V_1 V_2$, 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 which 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)

$$\widehat{OR}_{V_1, V_2, V_3, V_4} = \exp(\hat{\beta} + \hat{\delta}_1 V_1 + \hat{\delta}_2 V_2 + \hat{\delta}_4 V_4),$$

where $\hat{\delta}_1$, $\hat{\delta}_2$, and $\hat{\delta}_4$ are coefficients of EV_1 , EV_2 , and $EV_4 = EV_1 V_2$

$$\widehat{OR} = \exp(\hat{\beta} + \hat{\delta}_1 V_1 + \hat{\delta}_2 V_2 + \hat{\delta}_4 V_4)$$

\widehat{OR} differs for different specifications of V_1, V_2, V_4

gold standard \widehat{OR} ,

- controls for all potential confounders
- gives baseline \widehat{OR}

$$\widehat{OR}^* = \exp(\hat{\beta}^* + \hat{\delta}_1^* V_1 + \hat{\delta}_2^* V_2 + \hat{\delta}_4^* V_4),$$

where $\hat{\beta}^*$, $\hat{\delta}_1^*$, $\hat{\delta}_2^*$, $\hat{\delta}_4^*$ are coefficients in model without V_3

Model without V_3 :

$E, V_1, V_2, V_4, EV_1, EV_2, EV_4$

Model with V_3 :

$E, V_1, V_2, \textcircled{V}_3, V_4, EV_1, EV_2, EV_4$

Possible that

$$\hat{\beta} \neq \hat{\beta}^*, \hat{\delta}_1 \neq \hat{\delta}_1^*, \hat{\delta}_2 \neq \hat{\delta}_2^*, \hat{\delta}_4 \neq \hat{\delta}_4^*$$

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 .

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.

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.

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.

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.

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:

$$\widehat{OR} = \exp(\hat{\beta} + \hat{\delta}_1 V_1 + \hat{\delta}_2 V_2 + \hat{\delta}_4 V_4)$$

model without V_3 :

$$\widehat{OR}^* = \exp(\hat{\beta}^* + \hat{\delta}_1^* V_1 + \hat{\delta}_2^* V_2 + \hat{\delta}_4^* V_4)$$

$(\hat{\beta}, \hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_4)$ vs. $(\hat{\beta}^*, \hat{\delta}_1^*, \hat{\delta}_2^*, \hat{\delta}_4^*)$

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)

$$\widehat{OR} = \exp\left(\widehat{\beta} + \widehat{\delta}_1 V_1 + \widehat{\delta}_2 V_2 + \widehat{\delta}_4 V_4\right)$$

linear function

$\widehat{\beta}$, $\widehat{\delta}_1$, $\widehat{\delta}_2$, $\widehat{\delta}_4$ on log odds ratio scale;
but odds ratio scale is clinically relevant

Log odds ratio scale:

$$\widehat{\beta} = -12.69 \text{ vs. } \widehat{\beta}^* = -12.72$$

$$\widehat{\delta}_1 = 0.0692 \text{ vs. } \delta_1^* = 0.0696$$

Odds ratio scale:

$$\text{calculate } \widehat{OR} = \exp\left(\widehat{\beta} + \sum \delta_j W_j\right)$$

for different choices of W_j

Gold standard OR:

$$\widehat{OR} = \exp\left(\widehat{\beta} + \widehat{\delta}_1 V_1 + \widehat{\delta}_2 V_2 + \widehat{\delta}_4 V_4\right)$$

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{OR}	\widehat{OR}	\widehat{OR}
$V_2=200$	\widehat{OR}	\widehat{OR}	\widehat{OR}

Model without V_3 :

$$\widehat{OR}^* = \exp\left(\widehat{\beta}^* + \widehat{\delta}_1^* V_1 + \widehat{\delta}_2^* V_2 + \widehat{\delta}_4^* V_4\right)$$

	$V_1=20$	$V_1=30$	$V_1=40$
$V_2=100$	\widehat{OR}^*	\widehat{OR}^*	\widehat{OR}^*
$V_2=200$	\widehat{OR}^*	\widehat{OR}^*	\widehat{OR}^*

Compare tables of

\widehat{OR} s vs. \widehat{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 $\widehat{\beta}$ from -12.69 to -12.72 and a change in $\widehat{\delta}_1$ 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 which 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.

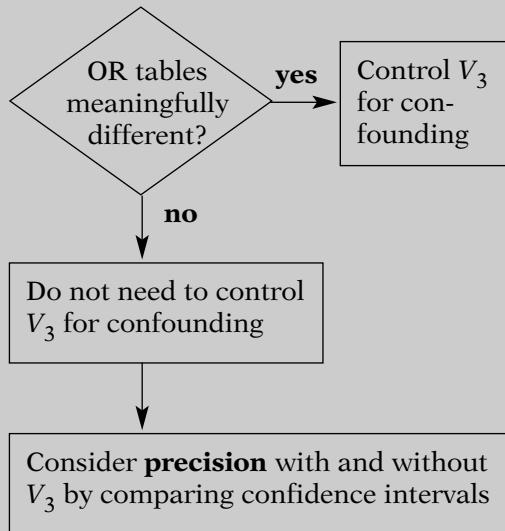
Thus, to assess whether we need to control for confounding from V_3 , we need to compare two tables of odds ratios, one for the gold standard and the other for the model which does not contain V_3 .

EXAMPLE (continued)

gold standard OR		
\widehat{OR}	\widehat{OR}	\widehat{OR}
\widehat{OR}	\widehat{OR}	\widehat{OR}

corresponding odds ratios

\widehat{OR}^* (excludes V_3)		
\widehat{OR}^*	\widehat{OR}^*	\widehat{OR}^*
\widehat{OR}^*	\widehat{OR}^*	\widehat{OR}^*



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.

Gain in precision?

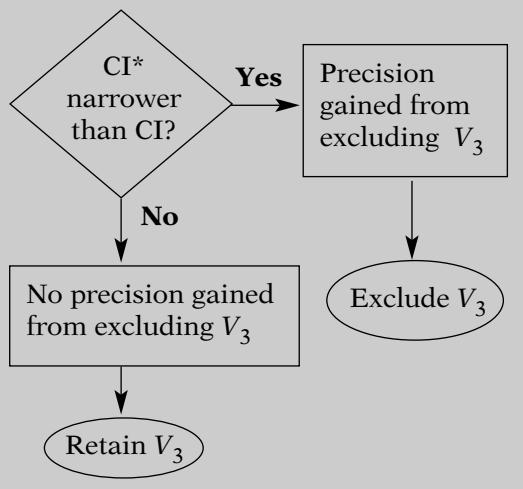
gold standard CI

CI	CI	CI
CI	CI	CI

 CI^* (excludes V_3)

CI^*	CI^*	CI^*
CI^*	CI^*	CI^*

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)

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 confounders

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

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

$$C's: \underbrace{\text{AGE}, \text{CHL}}_{\text{continuous}} \quad \underbrace{\text{SMK}, \text{ECG}, \text{HPT}}_{(0, 1)}$$

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

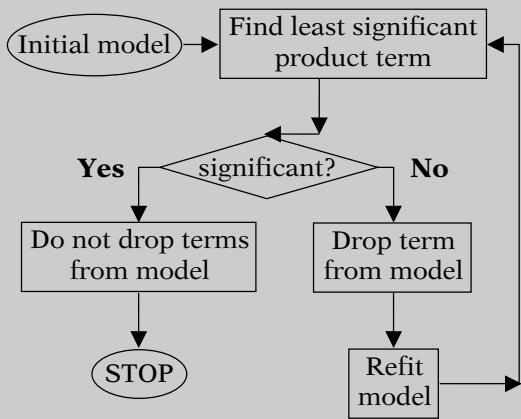
Highest order in model: EV_i

no EV_iV_j or V_iV_j terms

Next step:

interaction assessment using
backward elimination

Backward elimination:



Interaction results:

eliminated	remaining
$\text{CAT} \times \text{AGE}$	$\text{CAT} \times \text{CHL}$
$\text{CAT} \times \text{SMK}$	$\text{CAT} \times \text{HPT}$
$\text{CAT} \times \text{ECG}$	

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 which 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 procedure to remove variables. For interaction terms, we proceed by eliminating 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 backward procedure allows us to eliminate the product terms of $\text{CAT} \times \text{AGE}$, $\text{CAT} \times \text{SMK}$, and $\text{CAT} \times \text{ECG}$. The remaining interaction terms are $\text{CAT} \times \text{CHL}$ and $\text{CAT} \times \text{HPT}$.

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
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

V's *W's*
 $\text{CH} = \text{CAT} \times \text{HPT}$ and $\text{CC} = \text{CAT} \times \text{CHL}$
 remain in all further models

Confounding assessment:

Step 1. Variables in model:

CAT, AGE, CHL, SMK, ECG, HPT
V's
CAT × CHL, CAT × HPT,
EV's

All five *V*'s still in model after interaction

Hierarchy Principle:

- determine *V*'s that cannot be eliminated
- all lower-order components of significant product terms remain

$\text{CAT} \times \text{CHL}$ significant \Rightarrow CAT and CHL components

$\text{CAT} \times \text{HPT}$ significant \Rightarrow CAT and HPT components

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 \times HPT and CC equals CAT \times 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.

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 \times CHL and CAT \times HPT.

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.

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.

The two significant product terms in our model are CAT \times CHL and CAT \times HPT. The lower-order components of CAT \times CHL are CAT and CHL. The lower-order components of CAT \times HPT are CAT and HPT.

EXAMPLE (continued)

Thus, retain CAT, CHL, and HPT in all further models

Candidates for elimination:

AGE, SMK, ECG

Assessing confounding:

do coefficients in \widehat{OR} expression change?

$$\widehat{OR} = \exp(\hat{\beta} + \hat{\delta}_1 CHL + \hat{\delta}_2 HPT),$$

where

$\hat{\beta}$ = coefficient of CAT

$\hat{\delta}_1$ = coefficient of CC = CAT \times CHL

$\hat{\delta}_2$ = coefficient of CH = CAT \times HPT

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.

This leaves the V variables AGE, SMK, and ECG as still being candidates for elimination as possible nonconfounders.

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.

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.

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.

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.

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

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{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.

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

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 which deletes the three eligible variables AGE, SMK, and ECG.

Gold standard \widehat{OR} :

$$\widehat{OR} = \exp(-12.6894 + 0.0692CHL - 2.3318HPT)$$

	HPT = 0	HPT = 1
CHL = 200	$\widehat{OR} = 3.16$	$\widehat{OR} = 0.31$
CHL = 220	$\widehat{OR} = 12.61$	$\widehat{OR} = 1.22$
CHL = 240	$\widehat{OR} = 50.33$	$\widehat{OR} = 4.89$
CHL = 200, HPT = 0	$\Rightarrow \widehat{OR} = 3.16$	
CHL = 220, HPT = 1		$\Rightarrow \widehat{OR} = 1.22$

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 \times CHL and CAT \times 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)

$$\text{CHL} = 200, \text{HPT} = 0 \Rightarrow \widehat{\text{OR}} = (3.16)$$

$$\text{CHL} = 220, \text{HPT} = 1 \Rightarrow \widehat{\text{OR}} = (1.22)$$

$\widehat{\text{OR}}$ with AGE, SMK, ECG deleted:

$$\widehat{\text{OR}}^* = \exp(-12.7324 + 0.0712\text{CHL} - 2.2584\text{HPT})$$

	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	$\widehat{\text{OR}}^* = 79.11$	$\widehat{\text{OR}}^* = 8.34$

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

Cannot simultaneously drop AGE, SMK, and ECG from model

gold standard

other models

vs.

Other models: delete AGE and SMK or delete AGE and ECG, etc.

Result: cannot drop AGE, SMK, or ECG

Final model:

E: CAT

five V's: CHL, HPT, AGE, SMK, ECG

two interactions: CAT \times CHL, CAT \times HPT

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.

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.

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.

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.

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

Confounding and precision difficult if interaction (subjective)

Caution: do not sacrifice validity for minor gain in precision

Summary result for final model:

CHL	Table of \widehat{OR}		Table of 95% CIs	
	HPT=0	HPT=1	HPT=0	HPT=1
200	3.16	0.31	(0.89, 11.03)	(0.10, 0.91)
220	12.61	1.22	(3.65, 42.94)	(0.48, 3.10)
240	50.33	4.89	(11.79, 212.23)	(1.62, 14.52)

Use to draw meaningful conclusions

$$\text{CHL} \nearrow \Rightarrow \widehat{OR}_{\text{CAT}, \text{CHD}} \nearrow$$

$$\text{CHL fixed: } \widehat{OR}_{\text{CAT}, \text{CHD}}_{\text{HPT}=0} > \widehat{OR}_{\text{CAT}, \text{CHD}}_{\text{HPT}=1}$$

All CIs are wide

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.

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.

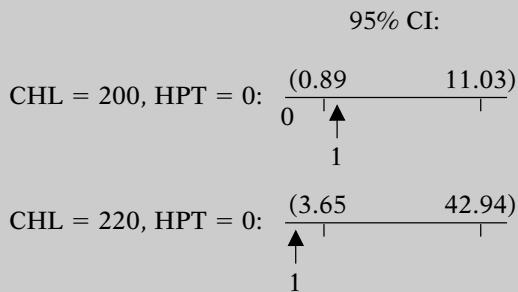
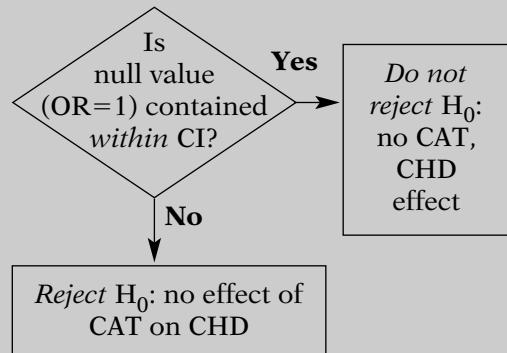
To conclude this example, we point out that, using the final model, a summary of the results of the analysis can be made in terms of the table of odds ratios and the corresponding table of confidence intervals.

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.

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:



Test results at 5% level:

- $\text{CHL} = 200, \text{HPT} = 0$: no significant CAT, CHD effect
- $\text{CHL} = 220, \text{HPT} = 0$: no significant CAT, CHD effect

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.

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*

One-tailed tests:
use large sample

$$Z = \frac{\text{estimate}}{\text{standard error}}$$

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.

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.

SUMMARY

Chapter 6

- overall guidelines for three stages
- focus: variable specification
- HWF model

Chapter 7

Focus: interaction and confounding assessment

Interaction: use hierarchical backward elimination

Use **Hierarchy Principle** to identify lower-order components that cannot be deleted (EV 's, V_i 's, and V_iV_j 's)

Confounding: *no* statistical testing:
Compare whether \widehat{OR} meaningfully changes when V 's are deleted

Drop nonconfounders if precision is gained by examining CIs

No interaction: assess confounding by monitoring changes in $\hat{\beta}$, the coefficient of E

A brief summary of this presentation is now given. This has been the second of two chapters on modeling strategy. In Chapter 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).

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.

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 (that is, terms of the form EV) and to lower-order terms involving potential confounders of the form V_i or V_iV_j .

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.

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.

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)

Interaction: Safe (for validity) to keep all V's in model

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.

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. Analysis of Matched Data

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 concerns the use of logistic modeling to assess matched data.

Detailed Outline

I. Overview (pages 194–195)

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, \dots, C_p .
 - ii. 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.
 - iii. 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.
- 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 EV_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 (195–198)

- 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 if chunk test is significant; testing procedure involves likelihood ratio statistic.
- D. Example.

III. Confounding and precision assessment when no interaction (pages 199–203)

- 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 V 's specified initially.
- C. Identify subsets of V 's 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 203–211)

- 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 211–218)

- 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 (McLawns 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

- 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:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta \text{HT} + \gamma_1 \text{HS} + \gamma_2 \text{CT} + \gamma_3 \text{AGE} + \gamma_4 \text{SEX} \\ + \delta_1 \text{HT} \times \text{AGE} + \delta_2 \text{HT} \times \text{SEX}.$$

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.

- Using the model given in Exercise 1, describe briefly how to carry out a backward elimination procedure to assess interaction.
- 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.)
- 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.
- 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.
- Suppose the interaction assessment stage finds that the interaction terms HT \times AGE and HT \times 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?
- 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?
- Describe briefly how you would assess confounding for the model considered in Exercises 6 and 7.
- Suppose that the variable SEX is determined to be a *nonconfounder*, whereas all other V variables in the model (of Exercise 1) need to be controlled. Describe briefly how you would assess whether the variable SEX needs to be controlled for precision reasons.
- 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:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta \text{SMK} + \sum \gamma_i^* V_i^* + \gamma_1 \text{NS} + \gamma_2 \text{AS} + \gamma_3 \text{NS} \times \text{AS} \\ + \delta_1 \text{SMK} \times \text{NS} + \delta_2 \text{SMK} \times \text{AS} + \delta_3 \text{SMK} \times \text{NS} \times \text{AS},$$

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 $\text{SMK} \times \text{NS}$. 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 $\text{SMK} \times \text{NS}$), 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 $\text{NS} \times \text{AS}$ 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 $\text{NS} \times \text{AS}$ should be retained in the model? (In answering this question, consider both confounding and precision issues.)

7. Suppose the variable $NS \times AS$ 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 \times 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 $HT \times AGE$ and $HT \times SEX$ 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 $\exp(\beta)$, 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 $\exp(\beta)$ changes as one or more of the V variables are dropped from the model. If, for example, HS and CT are dropped and $\exp(\beta)$ does not change from the (gold standard) model containing all V 's, 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 $\exp(\beta + \delta_1 \text{AGE} + \delta_2 \text{SEX})$.
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 SEX 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 SEX. Because the odds ratio is defined by the expression $\exp(\beta + \delta_1 \text{AGE} + \delta_2 \text{SEX})$, a table of confidence intervals for both the model without SEX and with SEX will need to be obtained by specifying different values for the effect modifiers AGE and SEX. To assess whether SEX needs to be controlled for precision reasons, one must compare these tables of confidence intervals. If the confidence intervals when SEX is not in the model are narrower in some overall sense than when SEX is in the model, precision is gained by dropping SEX. Otherwise, SEX should be controlled as precision is not gained when the SEX 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

Analysis of Matched Data Using Logistic Regression

Contents	
Introduction	228
Abbreviated Outline	228
Objectives	229
Presentation	230
Detailed Outline	253
Practice Exercises	257
Test	261
Answers to Practice Exercises	263

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. Finally, we illustrate the analysis of matched data using logistic regression with an application that involves matching as well as control variables not involved in matching.

Abbreviated Outline

The outline below gives the user a preview of this chapter. A detailed outline for review purposes follows the presentation.

- I. Overview (page 230)
- II. Basic features of matching (pages 230–232)
- III. Matched analyses using stratification (pages 232–235)
- IV. The logistic model for matched data (pages 235–238)
- V. An application (pages 238–241)
- VI. Assessing interaction involving matching variables (pages 242–244)
- VII. Pooling matching strata (pages 245–247)
- VIII. Analysis of matched follow-up data (pages 247–251)

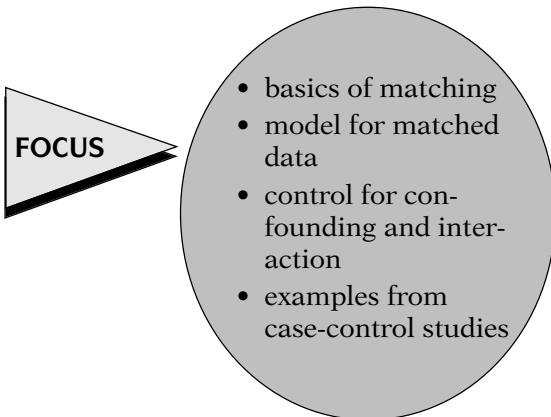
Objectives

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.

Presentation

I. Overview



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

Case-control study:



our focus referent = controls
 index = cases

Follow-up study:

referent = unexposed
index = exposed

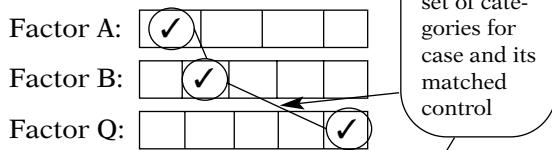
For example, if the matching factor is age, then matching on age would constrain the referent group to have essentially the same age structure as the index group.

In a case-control study, the referent group consists of the controls, which is compared to an index group of cases.

In a follow-up study, the referent group consists of unexposed subjects, which is compared to 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:



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., $R = 4 \rightarrow$ 4-to-1)	R -to-1

R may vary from case to case

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

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.

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.

As for whether to match or not in a given study, there are both advantages and disadvantages to consider.

To match or not to match

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*

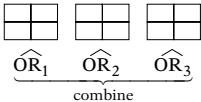
- to find matches
- information loss due to discarding controls

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

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.

Matching	No matching
Correct estimate?	
YES	YES
Apropriate analysis?	
YES ↓ MATCHED (STRATIFIED) ANALYSIS ↓ SEE SECTION III	YES ↓ STANDARD STRATIFIED ANALYSIS 30-39 40-49 50-59  combine

Validity is not an important reason for matching (validity: getting the right answer)

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.

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

The analysis of matched data can be carried out using a stratified analysis in which the strata consist of the collection of matched sets.

Strata = matched sets

Special case

Case-control study

100 matched pairs

$n = 200$

100 strata = 100 matched pairs

2 observations per stratum

1st pair

2nd pair

100th pair

$E \quad \bar{E}$		$E \quad \bar{E}$		$E \quad \bar{E}$	
D		D		D	
\bar{D}		\bar{D}		\bar{D}	

$E \quad \bar{E}$		$E \quad \bar{E}$		$E \quad \bar{E}$	
D		D		D	
\bar{D}		\bar{D}		\bar{D}	

$E \quad \bar{E}$		$E \quad \bar{E}$		$E \quad \bar{E}$	
D		D		D	
\bar{D}		\bar{D}		\bar{D}	

$E \quad \bar{E}$		$E \quad \bar{E}$		$E \quad \bar{E}$	
D		D		D	
\bar{D}		\bar{D}		\bar{D}	

Four possible forms:

$E \quad \bar{E}$		$E \quad \bar{E}$		$E \quad \bar{E}$	
D	1	0	1	W pairs	
\bar{D}	1	0	1		

$E \quad \bar{E}$		$E \quad \bar{E}$		$E \quad \bar{E}$	
D	1	0	1	X pairs	
\bar{D}	0	1	1		

$E \quad \bar{E}$		$E \quad \bar{E}$		$E \quad \bar{E}$	
D	0	1	1	Y pairs	
\bar{D}	1	0	1		

$E \quad \bar{E}$		$E \quad \bar{E}$		$E \quad \bar{E}$	
D	0	1	1	Z pairs	
\bar{D}	0	1	1		

$$W + X + Y + Z = \text{total number of pairs}$$

EXAMPLE

$$W = 30, X = 30, Y = 10, Z = 30$$

$$W + X + Y + Z = 30 + 30 + 10 + 30 = 100$$

Analysis: two equivalent ways

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.

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.

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.

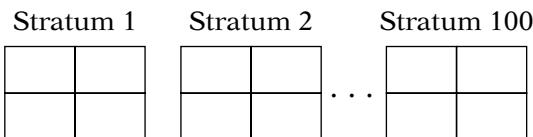
In the third table, the case is unexposed and the control is exposed.

And in the fourth table, both the case and the control are unexposed.

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.

For example, we may have W equals 30, X equals 30, Y equals 10, and Z equals 30, which sums to 100.

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 $\widehat{\text{MOR}}$

	D	\bar{E}
E	E	\bar{E}
\bar{E}	W	Y

D	$E \quad \bar{E}$	$D \quad E \quad \bar{E}$	$D \quad E \quad \bar{E}$	$D \quad E \quad \bar{E}$
\bar{D}	1 0 1 0	1 0 1 0	1 0 1 0	1 0 1 0
	W	X	Y	Z

$$\chi^2 = \frac{(X - Y)^2}{X + Y}, \text{ df} = 1$$

McNemar's test

McNemar's test = MH test for pair-matching

$$\widehat{\text{MOR}} = X/Y$$

EXAMPLE

	D	
E	E	\bar{E}
\bar{E}	$W=30$	$Y=10$

	D	
E	30	10
\bar{E}	30	30

	D	
E	30	10
\bar{E}	30	30

$$\chi^2 = \frac{(30 - 10)^2}{30 + 10} = \frac{400}{40} = 10.0$$

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. See Kleinbaum et al., *Epidemiologic Research*, (1982) for details.

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.

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.

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**.

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 Mantel–Haenszel odds ratio estimate can be calculated as X/Y .

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)

$\chi^2 \sim \text{chi square } 1 \text{ df}$
under H_0 : OR = 1

$P << 0.01$, significant

$$\widehat{\text{MOR}} = \frac{X}{Y} = 3$$

Analysis for R -to-1 and mixed matching
use stratified analysis

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.

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	\bar{E}	
D	1	0	1
\bar{D}	1	3	4
			5

R -to-1 or mixed matching

use χ^2 (MH) and $\widehat{\text{MOR}}$ for stratified data

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.

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

Advantage of modeling
can control for variables *other* than
matched variables

A third approach to carrying out the analysis of matched data involves logistic regression modeling.

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

Logistic model for matched data includes control of variables not matched

Stratified analysis inefficient:
data is discarded

Matched data:
use conditional ML estimation
(number of parameters large relative to n)

Pair-matching:

$$\widehat{\text{OR}}_U = (\widehat{\text{OR}}_C)^2$$

↑
overestimate

Principle

matched analysis \Rightarrow stratified analysis

- strata are matched sets, e.g., pairs
- strata defined using dummy (indicator) variables

$E = (0, 1)$ exposure

C_1, C_2, \dots, C_p control variables

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.

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.

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.

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.

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.

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.

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 C 's matched by design
- remaining C 's not matched

$D = (0, 1)$ disease

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

Some X 's: V_{1i} dummy variables (matched strata)

Some X 's: V_{2i} variables (potential confounders)

Some X 's: product terms EW_j
(Note: W 's usually V_2 's)

The model

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \sum \underbrace{\gamma_{1i} V_{1i}}_{\text{matching}} + \sum \underbrace{\gamma_{2i} V_{2i}}_{\text{confounders}} + E \sum \underbrace{\delta_j W_j}_{\text{interaction}}$$

EXAMPLE

Pair-matching by AGE, RACE, SEX

100 matched pairs

99 dummy variables

$$V_{1i} = \begin{cases} 1 & \text{if } i\text{th matched pair} \\ 0 & \text{otherwise} \end{cases} \quad i = 1, 2, \dots, 99$$

$$V_{11} = \begin{cases} 1 & \text{if first matched pair} \\ 0 & \text{otherwise} \end{cases}$$

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.

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 .

We also have a collection of X 's which are dummy variables to indicate the different matched strata; these variables are denoted as V_1 variables.

Further, we have a collection of X 's which are defined from the C 's not involved in the matching and represent potential confounders in addition to the matched variables. These potential confounders are denoted as V_2 variables.

And finally, we have a collection of X 's which are product terms of the form E times W , where the W 's denote potential interaction variables. Note that the W 's will usually be defined in terms of the V_2 variables.

The logistic model for matched analysis is then given in logit form as shown here. In this model, the γ_{1i} are coefficients of the dummy variables for the matching strata, the γ_{2i} are the coefficients of the potential confounders not involved in the matching, and the δ_j are the coefficients of the interaction variables.

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 i th 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,

EXAMPLE (continued)

$$V_{12} = \begin{cases} 1 & \text{if second matched pair} \\ 0 & \text{otherwise} \end{cases}$$

⋮

$$V_{1,99} = \begin{cases} 1 & \text{if 99th matched pair} \\ 0 & \text{otherwise} \end{cases}$$

1st matched set

$$V_{11} = 1, V_{12} = V_{13} = \dots = V_{1,99} = 0$$

99th matched set

$$V_{1,99} = 1, V_{11} = V_{12} = \dots = V_{1,98} = 0$$

100th matched set

$$V_{11} = V_{12} = \dots = V_{1,99} = 0$$

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.

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

$$\logit P(\mathbf{X}) = \alpha + \beta E + \sum \gamma_{1i} V_{1i} + \sum \gamma_{2i} V_{2i} + E \sum \delta_j W_j$$

$$ROR = \exp(\beta + \sum \delta_j W_j)$$

Note: two types of V variables are controlled

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 .

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

$$D = MI_{0,1}$$

$$E = SMK_{0,1}$$

$$C_1 = AGE, C_2 = RACE, C_3 = SEX, C_4 = HOSPITAL$$

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.

$$C_5 = SBP \quad C_6 = ECG$$

not matched

EXAMPLE (continued)

$n = 117$ (39 matched sets)

The model:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta \text{SMK} + \sum_{i=1}^{38} \gamma_{1i} V_{1i}$$

$$= \gamma_{21} \text{SBP} + \gamma_{22} \text{ECG}$$

↑ confounders

$$+ \text{SMK} (\delta_1 \text{SBP} + \delta_2 \text{ECG})$$

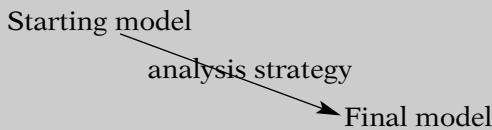
↑ modifiers

$$\text{ROR} = \exp(\beta + \delta_1 \text{SBP} + \delta_2 \text{ECG})$$

β = coefficient of E

δ_1 = coefficient of $E \times \text{SBP}$

δ_2 = coefficient of $E \times \text{ECG}$



Estimation method:

- ✓ conditional ML estimation
- (also, we illustrate unconditional ML estimation)

Interaction:
 $\text{SMK} \times \text{SBP}$ and $\text{SMK} \times \text{ECG}$?

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 logistic model for the above situation can be defined as follows: logit $P(\mathbf{X})$ equals α plus β times SMK plus the sum of 38 terms of the form γ_{1i} times V_{1i} , where V_{1i} 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.

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.

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 $E \times \text{SBP}$ and $E \times \text{ECG}$.

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.

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.

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 $\text{SMK} \times \text{SBP}$ and $\text{SMK} \times \text{ECG}$ are retained in the model.

EXAMPLE (continued)

Chunk test

$$H_0: \delta_1 = \delta_2 = 0$$

where

δ_1 = coefficient of SMK \times SBP

δ_2 = coefficient of SMK \times ECG

$$LR = (-2 \ln \hat{L}_R) - (-2 \ln \hat{L}_F)$$

R = reduced model F = full model
(no interaction) (interaction)

log likelihood statistics

$$-2 \ln \hat{L}$$

$$LR \sim \chi^2_2$$

Number of parameters tested = 2

$$-2 \ln \hat{L}_F = 60.23$$

$$-2 \ln \hat{L}_R = 60.63$$

$$LR = 60.63 - 60.23 = 0.40$$

$P > 0.10$ (no significant interaction)

Therefore, drop SMK \times SBP and
SMK \times ECG from model

Backward elimination: same conclusion

$$\text{logit } P(\mathbf{X}) = \alpha + \beta \text{SMK} + \sum \gamma_{1i} V_{1i} \\ + \gamma_{21} \text{SBP} + \gamma_{22} \text{ECG}$$

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.

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.

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.

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.

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.

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.

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)

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

$V's$ in model	$\text{OR} = e^{\beta}$	95% CI
SBP and ECG	$C \underline{2.07}$	(0.69, 6.23)
	$U \underline{3.38}$	
SBP only	$C \underline{2.08}$	(0.72, 6.00)
	$U \underline{3.39}$	
ECG only	$C \underline{2.05}$	(0.77, 5.49)
	$U \underline{3.05}$	
Neither	$C \underline{2.32}$	(0.93, 5.79)
	$U \underline{3.71}$	

C = conditional estimate

U = unconditional estimate

Minimal confounding:

gold standard $\widehat{\text{OR}} = 2.07$, essentially same as other $\widehat{\text{OR}}$

But 2.07 moderately different from 2.32, so we control for *at least* one of SBP and ECG

Narrowest CI: control for ECG only

Most precise estimate:
control for ECG only

All CI are wide and include 1

Overall conclusion:
adjusted $\widehat{\text{OR}} \approx 2$, but is nonsignificant

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.

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.

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.

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.

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.

Thus, the most precise estimate of the effect is obtained when ECG is controlled, along, of course, with the matching variables.

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

Interaction terms:

$\text{SMK} \times \text{SBP}$, $\text{SMK} \times \text{ECG}$

tested using LR test

Interaction between

SMK and matching variables?

Two options.

Option 1

Add product terms of the form

$$E \times V_{1i}$$

$$\begin{aligned} \text{logit } P(X) &= \alpha + \beta E + \sum_i \gamma_{1i} V_{1i} + \sum_j \gamma_{2j} V_{2j} \\ &\quad + E \sum_i \delta_{1i} V_{1i} + E \sum_k \delta_k W_k \end{aligned}$$

where

V_{1i} = dummy variables for matching strata

V_{2j} = other covariates (not matched)

W_k = effect modifiers defined from other covariates

The previous section considered a study of the relationship between smoking (SMK) and myocardial 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.

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.

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.

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.

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 \Rightarrow No interaction involving matching variables

Significant \Rightarrow Interaction involving matching variables

\Rightarrow Carry out backward elimination of $E \times V_{1i}$ terms

Criticisms of Option 1

- Difficult to determine which of several matching variables are effect modifiers. (The V_{1i} represent matching strata, not matching variables.)
- Not enough data to assess interaction (number of parameters may exceed n)

Option 2

Add product terms of the form

$$E \times W_{1m}$$

where W_{1m} are *matching variables*

$$\begin{aligned} \text{logit } P(\mathbf{X}) = \alpha + \beta E &= \sum_i \gamma_{1i} V_{1i} + \sum_j \gamma_{2j} V_{2j} \\ &+ E \sum_m \delta_{1m} W_{1m} + E \sum_k \delta_{2k} W_{2k} \end{aligned}$$

where

W_{1m} = matching variables in original form

W_{2k} = effect modifiers defined from other covariates (not matched)

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)

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.

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.

A second option for assessing interaction involving matching variables is to consider product terms of the form $E \times W_{1m}$, where the W_{1m} 's are the actual matching variables.

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 $E \times W_{1m}$ involving the matching variables, and $E \times W_{2k}$ terms, where the W_{2k} are effect modifiers defined from the unmatched covariates.

EXAMPLE (continued)

Option 2

Test H_0 : All $\delta_{1m} = 0$.
(Chunk test)

Not significant \Rightarrow No interaction involving matching variables

Significant \Rightarrow Interaction involving matching variables
 \Rightarrow Carry out Backwards Elimination of $E \times W_{1m}$ terms

Criticism of Option 2

The model is technically not HWF.

$E \times W_{1m}$ in model but not W_{1m}

(Also, V_{1j} in model but not $E \times V_{1i}$)

	<u>Option 1</u>	<u>Option 2</u>
Interpretable?	No	Yes
HWF?	Yes	No (but almost yes)

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.

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.

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 Chapter 6 for a discussion of the HWF criterion.)

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.

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?

Case-control study:

- pair-match on SMK (ever versus never)
- 100 cases (i.e., $n = 200$)
- Smokers—60 matched pairs
- Nonsmokers—40 matched pairs

Matched pair A Matched pair B

Case A-Smoker Case B-Smoker

Control A-Smoker \leftrightarrow Control B-Smoker
(interchangeable)

Controls for matched pairs A and B
are interchangeable



Matched pairs A and B
are **exchangeable**
(definition)

Smokers: 60 matched pairs are
exchangeable

Nonsmokers: 40 matched pairs are
exchangeable

Ignoring exchangeability



Use stratified analysis with
100 strata, e.g., McNemar's test

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.

Suppose smoking status (SMK), defined as ever versus 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.

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.

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.

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.

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 Section III.

Ignore exchangeability? **No!!!**

Treating such strata separately is artificial,

i.e., exchangeable strata are not unique

Analysis? Pool exchangeable matched sets

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.]

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 ($n = 40 \times 2$)

Matching on several variables



May be only a few exchangeable matched sets



Pooling has negligible effect on odds ratio estimates

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 non-smoker's stratum.

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 to an unpooled matched analysis.

However, pooling may greatly reduce the number of strata to be analyzed (e.g., from 100 to 2 strata)

If # of strata greatly reduced by pooling



Unconditional ML may be used if "appropriate"

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.

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

$OR_{\text{unconditional}}$ ***unbiased***

and

$CI_{\text{unconditional}}$ ***narrower than***
 $CI_{\text{conditional}}$

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

Unexposed and exposed groups have same distribution of matching variables.

	Exposed	Unexposed
White male	30%	30%
White female	20%	20%
Nonwhite male	15%	15%
Nonwhite female	35%	35%

Individual matching
or

Frequency matching (more convenient,
larger sample size)

Thus far we have considered only matched case-control data. We now focus on the analysis of matched cohort data.

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.

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.

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 # of strata)



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

P pairs
concordant

Type 2

	E	\bar{E}
D	1	0
\bar{D}	0	1

Q pairs
discordant

Type 3

	E	\bar{E}
D	0	1
\bar{D}	1	0

R pairs
discordant

Type 4

	E	\bar{E}
D	0	0
\bar{D}	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

		\bar{E}	
	D	P	Q
E	\bar{D}	R	S

Without pooling → McNemar's table

$$\widehat{\text{MRR}} = \frac{P+Q}{P+R} \quad \widehat{\text{MOR}} = \frac{Q}{R}$$

$$\chi^2_{\text{MH}} = \frac{(Q-R)^2}{Q+R}$$

$\widehat{\text{MOR}}$ and χ^2_{MH} use discordant pairs information

$\widehat{\text{MRR}}$ uses discordant and concordant pairs information

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.

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.

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.

As an example, consider a pair-matched follow-up study with 4830 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

Pair-matched follow-up study 4830 matched pairs

$E = \text{VS}$ (0=no, 1=yes)

$D = \text{MI}$ (0=no, 1=yes)

Matching variables: AGE and YEAR

EXAMPLE (continued)

McNemar's table:

		VS = 0	
	MI = 1	MI = 1	MI = 0
VS = 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 $\widehat{\text{MRR}} = \widehat{\text{MOR}} = Q/R = 20/16$, or 1.25. Notice that since $P = 0$ in this table, the MRR equals the 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 4790 discordant pairs not used
- Pooling exchangeable matched sets more appropriate analysis
- Frequency matching more appropriate than individual matching

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 4790 concordant pairs was not needed, other than to distinguish such pairs from concordant pairs.

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.

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



Use logistic regression

No interaction model:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta VS + \sum_i \gamma_{1i} V_{1i} \\ + \gamma_{21} OBS + \gamma_{22} SMK$$

4830 total pairs \leftrightarrow 36 discordant pairs

same results

Need only analyze discordant pairs

Pair-matched case-control studies:

Use only discordant pairs

provided

no other control variables other than
matching variables

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.

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).

It turns out (from statistical theory) that the results from fitting the above model would be identical regardless of whether all 4380 matched pairs or just the 36 discordant matched pairs are input as the data.

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.

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 in order to obtain correct results.

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

•
•
•

- ✓ 8. Analysis of Matched Data
9. Polytomous Logistic Regression
 10. Ordinal Logistic Regression

The reader may wish 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 230)

Focus

- basics of matching
- model for matched data
- control of confounding and interaction
- examples

II. Basic features of matching (pages 230–232)

- 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 232–235)

- 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 235–238)

- 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:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \sum \gamma_{1i} V_{1i} + \sum \gamma_{2i} V_{2i} + E \sum \delta_j W_j$$

where V_{1i} are dummy variables identifying matched strata, V_{2i} are potential confounders based on variables not involved in the matching, and W_j 's are effect modifiers (usually) based on variables not involved in the matching.

- D. Odds ratio expression if E is coded as (0, 1):

$$\text{ROR} = \exp(\beta + \sum \delta_j W_j)$$

V. An application (pages 238–241)

- 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: 2.14,
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 242–244)

- A. Option 1: Add product terms of the form $E \times V_{1i}$, where V_{1i} are dummy variables for matching strata.

Model: $\text{logit } P(\mathbf{X}) = \alpha + \beta E + \sum \gamma_{1i} V_{1i} + \sum \gamma_{2j} V_{2j} + E \sum \delta_{1i} V_{1i} + E \sum \delta_{2j} W_k$

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 $E \times W_{1m}$, where W_{1m} are the matching variables in original form.

Model: $\text{logit } P(\mathbf{X}) = \alpha + \beta E + \sum \gamma_{1i} V_{1i} + \sum \gamma_{2j} V_{2j} + E \sum \delta_{1i} W_{1m} + E \sum \delta_{2k} W_{2k}$

where V_{2j} are other covariates (not matched) and W_{2k} are effect modifiers defined from other covariates.

Criticism of option 2:

- model is not HWF (i.e., $E \times W_{1m}$ in model but not W_{1m})
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 245–247)

- 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 247–251)

- 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:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta E + \sum \gamma_{1i} V_{1i} + \sum \gamma_{2i} V_{2i} + E \sum \delta_k W_k$$

- 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 $\widehat{MRR} = (P+Q)/(P+R)$, $\widehat{MOR} = Q/R$, and $\chi^2_{MH} = (Q - R)^2/(Q + R)$.
- F. Example: pair-matched follow-up study with 4830 matched pairs, $E=VS$ (vasectomy status), $D=MI$ (myocardial infarction status), match on AGE and YEAR (of follow-up); $P=0$, $Q=20$, $R=16$, $S=4790$.

$$\widehat{MRR} = 1.25 = \widehat{MOR}, \chi^2_{MH} = 0.44 \text{ (N.S.)}$$

Criticisms:

- information on 4790 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.

**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.

D	E	not D	
		1	2
not D	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.

D	E	SMK=1		2
		1	1	
not D	not E	0	2	4
SMK=0		SMK=0		3
D	E	2	1	
		2	1	
not D		not D		6
not D	not E	2	1	

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 non-smoker control of stratum 4 from the data set provided for Exercises 19–22. Let W = # concordant ($E=1, E=1$) pairs, X = # discordant ($E=1, E=0$) pairs, Y = # discordant ($E=0, E=1$) pairs, and Z = # 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 non-veteran = 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:

Variable	β	s_β	P -value	OR	95% CI for 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.

The following questions consider information 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:

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

11. What method of estimation was used to obtain estimates given in the above printout? Explain.
12. Why are the variables age and socioeconomic status missing from the printout given above, even though these were variables matched on in the study design?
13. State the logit form of the model used in the above printout.
14. Based on the printout above, is the product term $SMK \times NS$ significant? Explain.
15. Using the printout, 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.

16. Use the formula computed in Question 15 to compute numerical values for the estimated odds ratios when NS = 1 and NS = 0.
17. When NS = 1, the 95% confidence interval for the adjusted odds ratio for the effect of smoking on cervical cancer status is given by the limits (0.96, 5.44). Use this result and your estimate from Question 16 for NS = 1 to draw conclusions about the effect of smoking on cervical cancer status when NS = 1.
18. The following printout results from fitting a no interaction model to the cervical cancer data:

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

Based on this printout, compute the odds ratio for the effect of smoking, test its significance, and derive a 95% confidence interval of the odds ratio. Based on these results, what do you conclude about the effect of smoking on cervical cancer status?

Answers to Practice Exercises

1. F: cases are selected first, and controls are matched to cases
2. F: the age distribution for exposed persons is constrained to be the same as for unexposed 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 printout, the estimated odds ratio is $\exp(0.693)$, which equals 1.9997. The $\widehat{\text{MOR}}$ is computed as X/Y equals 40/20 equals 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

$$\exp\left[\hat{\beta} \pm 1.96 \sqrt{\widehat{\text{var}}(\hat{\beta})}\right]$$

which is computed to be

$$\exp(0.693 \pm 1.96 \times 0.274) = \exp(0.693 \pm 0.53704)$$

$$\text{which equals } (e^{0.15596}, e^{1.23004}) = (1.17, 3.42).$$

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. $\widehat{\text{ROR}} = \exp(\hat{\beta} + \hat{\delta} \text{PAR}).$

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. $\text{mOR(unpooled)} = \text{undefined}$; $\text{mOR(pooled)} = 2.5$
 30. Only choice a is correct.

9

Polytomous Logistic Regression

Contents	
Introduction	268
Abbreviated Outline	268
Objectives	269
Presentation	270
Detailed Outline	293
Practice Exercises	295
Test	297
Answers to Practice Exercises	298

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.

- I. Overview (pages 270–271)
- II. Polytomous logistic regression: An example with three categories (pages 272–275)
- III. Odds ratio with three categories (pages 275–279)
- IV. Statistical inference with three categories (pages 279–282)
- V. Extending the polytomous model to G outcomes and p predictors (pages 282–287)
- VI. Likelihood function for polytomous model (pages 288–290)
- VII. Polytomous versus multiple standard logistic regressions (page 291)

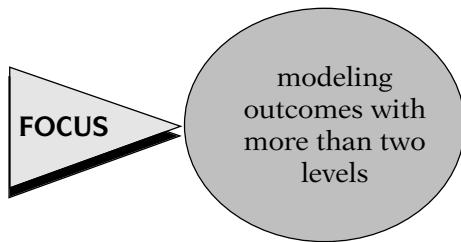
Objectives

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

One approach: dichotomize outcome

Change

0	1	2
---	---	---

to

0	1	2
---	---	---

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 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 versus none?
moderate versus mild?)

Alternate approach: use model for a
polytomous outcome

Nominal or ordinal outcome?

Nominal: different categories; no ordering

EXAMPLE

Endometrial cancer subtypes:

- adenosquamous
- adenocarcinoma
- other

Ordinal: levels have natural ordering

EXAMPLE

Tumor grade:

- well differentiated
- moderately differentiated
- poorly differentiated

Nominal outcome \Rightarrow Polytomous model

Ordinal outcome \Rightarrow Ordinal model or
polytomous model

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 versus none or moderate versus mild. This loss of detail may, in turn, affect the conclusions made about the exposure–disease relationship.

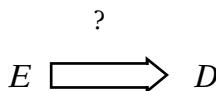
The detail of the original data coding can be retained through the use of models developed specifically for polytomous outcomes. The specific form that the model takes depends, in part, on whether the multi-level outcome variable is measured on a nominal or an ordinal scale.

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 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.

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 Chapter 10.

II. Polytomous Logistic Regression: An Example with Three Categories



EXAMPLE

Simplest case of polytomous model:

- outcome with three categories
- one dichotomous exposure variable

Data source:

Black/White Cancer Survival Study

$$E = \text{AGE} \begin{cases} 0 & \text{if 50--64} \\ 1 & \text{if 65--79} \end{cases}$$

$$D = \text{SUBTYPE} \begin{cases} 0 & \text{if Adenocarcinoma} \\ 1 & \text{if Adenosquamous} \\ 2 & \text{if Other} \end{cases}$$

SUBTYPE (0, 1, 2) uses arbitrary coding.

		AGE	
		50–64	65–79
		$E=0$	$E=1$
Adenocarcinoma $D=0$		77	109
Adenosquamous $D=1$		11	34
Other $D=2$		18	39

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)?

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.

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 on histological subtype among women with primary endometrial cancer. AGE, the exposure variable, is coded as 0 for ages 50–64 or 1 for ages 65–79. The disease variable, histological subtype, is coded 0 for adenocarcinoma, 1 for adenosquamous, and 2 for other.

There is no inherent order in the outcome variable. The 0, 1, and 2 coding of the disease categories is arbitrary.

The 3×2 table of the data is presented on the left.

Outcome categories:

A B C D

Reference (arbitrary choice)

Then compare:

A versus C, B vs. C, and D vs. C

EXAMPLE (continued)

Reference group = Adenocarcinoma

Two comparisons:

1. Adenosquamous (D=1)
versus Adenocarcinoma (D=0)
2. Other (D=2)
versus Adenocarcinoma (D=0)

Using data from table:

$$\widehat{OR}_{1vs0} = \frac{77 \times 34}{109 \times 11} = 2.18$$

$$\widehat{OR}_{2vs0} = \frac{77 \times 39}{109 \times 18} = 1.53$$

Dichotomous versus polytomous model:

Odds versus “odds-like” expressions

$$\text{logit } P(\mathbf{X}) = \ln \left[\frac{P(D=1 | \mathbf{X})}{P(D=0 | \mathbf{X})} \right]$$

$$= \alpha + \sum_{i=1}^p \beta_i X_i$$

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.

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.

Recall that for a dichotomous outcome variable coded as 0 or 1, the logit form of the logistic model, $\text{logit } P(\mathbf{X})$, is defined as the natural log of the odds for developing a disease for a person with a set of independent variables specified by \mathbf{X} . This logit form can be written as the linear function shown on the left.

Odds of disease: a ratio of probabilities

Dichotomous outcome:

$$\text{odds} = \frac{P(D=1)}{1-P(D=1)} = \frac{P(D=1)}{P(D=0)}$$

Polytomous outcome (three categories):

Use “odds-like” expressions for two comparisons



$$(1) \frac{P(D=1)}{P(D=0)} \quad (2) \frac{P(D=2)}{P(D=0)}$$

The logit form of model uses \ln of “odds-like” expressions

$$(1) \ln\left[\frac{P(D=1)}{P(D=0)}\right] \quad (2) \ln\left[\frac{P(D=2)}{P(D=0)}\right]$$

$$P(D=0) + P(D=1) + P(D=2) = 1$$

BUT

$$P(D=1) + P(D=0) \neq 1$$

$$P(D=2) + P(D=0) \neq 1$$

Therefore:

$$\frac{P(D=1)}{P(D=0)} \text{ and } \frac{P(D=2)}{P(D=0)}$$

“odds-like” but not true odds
(unless analysis restricted to two categories)

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 equals 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.

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.

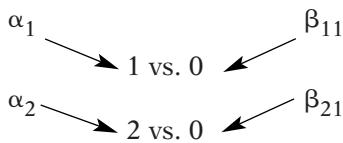
In polytomous logistic regression with three levels, we therefore define our model using two expressions for the natural log of these “odds-like” quantities. The first is the natural log of the probability that the outcome is in category 1 divided by the probability that the outcome is in category 0; 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.

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):

$$\ln \left[\frac{P(D=1|X_1)}{P(D=0|X_1)} \right] = \alpha_1 + \beta_{11} X_1$$

$$\ln \left[\frac{P(D=2|X_1)}{P(D=0|X_1)} \right] = \alpha_2 + \beta_{21} X_1$$



III. Odds Ratio with Three Categories

$$\left. \begin{array}{cc} \hat{\alpha}_1 & \hat{\alpha}_2 \\ \hat{\beta}_{11} & \hat{\beta}_{21} \end{array} \right\} \text{Estimates obtained as in SLR}$$

Special Case for One Predictor where $X_1 = 1$ or $X_1 = 0$

Because our example has three outcome categories and one predictor (i.e., AGE), 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 .

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 .

Both the alpha and beta terms have a subscript to indicate which comparison is being made (i.e., category **1** versus 0 or category **2** versus 0).

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).

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 versus category 0)
(Adenosquamous versus Adenocarcinoma)

OR_2 (category 2 versus category 0)
(Other versus Adenocarcinoma)

$$OR_1 = \frac{[P(D=1|X=1) / P(D=0|X=1)]}{[P(D=1|X=0) / P(D=0|X=0)]}$$

$$OR_2 = \frac{[P(D=2|X=1) / P(D=0|X=1)]}{[P(D=2|X=0) / P(D=0|X=0)]}$$

Adenosquamous versus Adenocarcinoma

$$OR_1 = \frac{\exp[\alpha_1 + \beta_{11}(1)]}{\exp[\alpha_1 + \beta_{11}(0)]} = e^{\beta_{11}}$$

Other versus Adenocarcinoma

$$OR_2 = \frac{\exp[\alpha_2 + \beta_{21}(1)]}{\exp[\alpha_2 + \beta_{21}(0)]} = e^{\beta_{21}}$$

$$OR_1 = e^{\beta_{11}} \quad OR_2 = e^{\beta_{21}}$$

They are different!

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.

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.

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 versus Adenocarcinoma) is e to the β_{11} .

The odds ratio for the second comparison (Other versus Adenocarcinoma) is e to the β_{21} .

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

$$\text{OR}_g = \exp[\beta_{g1} (X_1^{**} - X_1^*)] \text{ where } g=1,2$$

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		$\hat{\beta}_{11}$

EXAMPLE

Variable	Estimate	S.E.	Symbol
Intercept 1	-1.4534	0.2618	$\hat{\alpha}_2$
Intercept 2	-1.9459	0.3223	$\hat{\alpha}_1$
AGE	0.4256	0.3215	$\hat{\beta}_{21}$
AGE	0.7809	0.3775	$\hat{\beta}_{11}$

The special case of a dichotomous predictor can be generalized to include categorical or continuous predictors. To compare any two levels ($X_1 = X_1^{**}$ versus $X_1 = X_1^*$) 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 to the reference category (0).

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.

The results for the polytomous model examining histological subtype and age are presented on the left. The results were obtained from running PROC CATMOD in SAS.

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 versus Adenocarcinoma:

$$\widehat{\ln} \left[\frac{P(D = 2 | X_1)}{P(D = 0 | X_1)} \right] = -1.4534 + (0.4256)AGE$$

$$\widehat{OR}_2 = \exp[\widehat{\beta}_{21}] = \exp(0.4256) = 1.53$$

Adenosquamous versus Adenocarcinoma:

$$\widehat{\ln} \left[\frac{P(D = 1 | X_1)}{P(D = 0 | X_1)} \right] = -1.9459 + (0.7809)AGE$$

$$OR_1 = \exp[\widehat{\beta}_{11}] = \exp(0.7809) = 2.18$$

Special case

One dichotomous exposure \Rightarrow

polytomous model ORs = crude ORs

Interpretation of ORs

For older versus younger subjects:

- Other tumor category more likely than Adenocarcinoma ($\widehat{OR}_2 = 1.53$)
- Adenosquamous even more likely than Adenocarcinoma ($\widehat{OR}_1 = 2.18$)

The equation for the estimated log odds of Other (category 2) versus 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) versus Adenocarcinoma (category 0) is negative 1.9459 plus 0.7809 times age group.

Exponentiating the beta estimate for AGE 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 (ages 65–79) relative to younger subjects (ages 50–64) were more likely to have their tumors categorized as Other than as Adenocarcinoma ($\widehat{OR}_2 = 1.53$) and were even more likely to have their tumors classified as Adenosquamous than as Adenocarcinoma ($\widehat{OR}_1 = 2.18$).

Interpretation of alphas

Log odds where all X s 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 p). 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 in logistic regression. 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
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., standard logistic regression (SLR)].

95% CI for OR (one predictor)

$$\exp\left\{\hat{\beta}_{g1}(X_1^{**} - X_1^*) \pm 1.96(X_1^{**} - X_1^*)s_{\hat{\beta}_{g1}}\right\}$$

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.

Continuing with the endometrial cancer example, the estimated standard errors for the parameter estimates for AGE are 0.3215 for $\hat{\beta}_{21}$ and 0.3775 for $\hat{\beta}_{11}$.

EXAMPLE

Estimated standard errors:

$$s_{\hat{\beta}_{21}} = 0.3215 \quad s_{\hat{\beta}_{11}} = 0.3775$$

EXAMPLE (continued)

$$\begin{aligned}95\% \text{ CI for } OR_2 \\= \exp[0.4256 \pm 1.96(0.3215)] \\= (0.82, 2.87)\end{aligned}$$

$$\begin{aligned}95\% \text{ CI for } OR_1 \\= \exp[0.7809 \pm 1.96(0.3775)] \\= (1.04, 4.58)\end{aligned}$$

Likelihood ratio test

Assess significance of X_1

2β s tested at the same time
 \Downarrow

2 degrees of freedom

The 95% confidence interval for OR_2 is calculated as shown on the left as 0.82 to 2.87. The 95% confidence interval for OR_1 is calculated as 1.04 to 4.58.

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., $D=2$ versus $D=0$ and $D=1$ versus $D=0$). This affects the number of parameters tested and, therefore, the degrees of freedom associated with the test.

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

3 levels of D and 1 predictor
 \Downarrow
2 α s and 2 β s

Full model:

$$\ln \left[\frac{P(D=g|X_1)}{P(D=0|X_1)} \right] = \alpha_g + \beta_{g1} X_1, \quad g=1,2$$

Reduced model:

$$\ln \left[\frac{P(D=g)}{P(D=0)} \right] = \alpha_g, \quad g=1,2$$

$$H_0: \beta_{11} = \beta_{21} = 0$$

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

EXAMPLE

<u>$-2 \ln L$</u>	
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 AGE 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

$$H_0: \beta_{11} = 0 \text{ (category 1 vs. 0)}$$

$$Z = \frac{0.7809}{0.3775} = 2.07; \quad P = 0.04$$

Continuing with our example, the null hypothesis for the Adenosquamous versus 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 versus 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.

$$H_0: \beta_{21} = 0 \text{ (category 2 vs. 0)}$$

$$Z = \frac{0.4256}{0.3215} = 1.32, \quad P = 0.19$$

Conclusion: Is AGE significant?

⇒ Yes: Adenocarcinoma versus Adenosquamous

⇒ No: Other versus Adenosquamous.

Decision: retain or drop both β_{11} and β_{21} from model

At the 0.05 level of significance, we reject the null hypothesis for β_{11} but not for β_{21} . We conclude that AGE is statistically significant for the Adenosquamous versus Adenocarcinoma comparison (category 1 vs. 0), but not for the Other versus Adenocarcinoma comparison (category 2 vs. 0).

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 p Predictors

Adding More Independent Variables

Expanding the model to add more independent variables is straightforward. We can add p independent variables for each of the outcome comparisons.

$$\ln \left[\frac{P(D=1 | \mathbf{X})}{P(D=0 | \mathbf{X})} \right] = \alpha_1 + \sum_{i=1}^p \beta_{1i} X_i$$

$$\ln \left[\frac{P(D=2 | \mathbf{X})}{P(D=0 | \mathbf{X})} \right] = \alpha_2 + \sum_{i=1}^p \beta_{2i} X_i$$

Same procedures for OR, CI, and hypothesis testing

The log odds comparing category 1 to category 0 is equal to α_1 plus the summation of the p 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 p independent variables times their β_2 coefficients.

The procedures for calculation of the odds ratios, confidence intervals, and for hypothesis testing remain the same.

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 AGE on histological subtype ($D = 0, 1, 2$). The model now contains three predictor variables: $X_1 = \text{AGE}$, $X_2 = \text{ESTROGEN}$, and $X_3 = \text{SMOKING}$.

EXAMPLE

$$D = \text{SUBTYPE} \begin{cases} 0 & \text{if Adenocarcinoma} \\ 1 & \text{if Adenosquamous} \\ 2 & \text{if Other} \end{cases}$$

Predictors

$$X_1 = \text{AGE}$$

$$X_2 = \text{ESTROGEN}$$

$$X_3 = \text{SMOKING}$$

EXAMPLE (continued)

$$X_1 = \text{AGE} \quad \begin{cases} 0 & \text{if 50--64} \\ 1 & \text{if 65--79} \end{cases}$$

$$X_2 = \text{ESTROGEN} \quad \begin{cases} 0 & \text{if never user} \\ 1 & \text{if ever user} \end{cases}$$

$$X_3 = \text{SMOKING} \quad \begin{cases} 0 & \text{if former or never smoker} \\ 1 & \text{if current smoker} \end{cases}$$

Adenosquamous versus Adenocarcinoma

$$\ln \left[\frac{P(D=1 | X)}{P(D=0 | X)} \right] = \alpha_1 + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3$$

Other versus Adenocarcinoma

$$\ln \left[\frac{P(D=2 | X)}{P(D=0 | X)} \right] = \alpha_2 + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3$$

Recall that AGE is coded as 0 for ages 50–64 or 1 for ages 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.

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 .

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 .

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.

Variable	Estimate	S.E.	Symbol
Intercept 1	-1.2032	0.3190	$\hat{\alpha}_2$
Intercept 2	-1.8822	0.4025	$\hat{\alpha}_1$
AGE	0.2823	0.3280	$\hat{\beta}_{21}$
AGE	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	$\hat{\beta}_{13}$

EXAMPLE (continued)

Adenosquamous versus Adenocarcinoma

$$\widehat{OR}_1 = \frac{\exp[\hat{\alpha}_1 + \hat{\beta}_{11}(1) + \hat{\beta}_{12}(X_2) + \hat{\beta}_{13}(X_3)]}{\exp[\hat{\alpha}_1 + \hat{\beta}_{11}(0) + \hat{\beta}_{12}(X_2) + \hat{\beta}_{13}(X_3)]}$$

$$= \exp \hat{\beta}_{11} = \exp(0.9871) = 2.68$$

Other versus Adenocarcinoma

$$\widehat{OR}_2 = \frac{\exp[\hat{\alpha}_2 + \hat{\beta}_{21}(1) + \hat{\beta}_{22}(X_2) + \hat{\beta}_{23}(X_3)]}{\exp[\hat{\alpha}_2 + \hat{\beta}_{21}(0) + \hat{\beta}_{22}(X_2) + \hat{\beta}_{23}(X_3)]}$$

$$= \exp \hat{\beta}_{21} = \exp(0.2823) = 1.33$$

Interpretation of ORs

Three-variable versus one-variable model

Three-variable model

\Rightarrow AGE | ESTROGEN, SMOKING

One-variable model:

\Rightarrow AGE | no control variables

Odds ratios for effect of AGE:

Comparison	Model		
	AGE	ESTROGEN	SMOKING
	AGE		
1 vs. 0	2.68		2.18
2 vs. 0	1.33		1.53

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.

Suppose we are interested in the effect of AGE, controlling for the effects of ESTROGEN and SMOKING. The odds ratio for the effect of AGE 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 AGE 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.

Our interpretation of the results for the three-variable model differs from that of the one-variable model. The effect of AGE on the outcome is now estimated while controlling for the effects of ESTROGEN and SMOKING.

If we compare the model with three predictor variables with the model with only AGE included, the effect of AGE in the reduced model is weaker for the comparison of Adenosquamous to Adenocarcinoma ($\widehat{OR} = 2.18$ vs. 2.68), but is stronger for the comparison of Other to Adenocarcinoma ($\widehat{OR} = 1.53$ vs. 1.33).

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

$$= \exp[0.2832 \pm 1.96(0.3280)] \\ = (0.70, 2.52)$$

Likelihood ratio test
Wald tests

} same procedures
as with one predictor

Likelihood ratio test

reduced:	$\frac{-2 \ln L}{500.97}$
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 $\hat{\beta}_{21}$ 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 AGE significant?*

- ⇒ Yes: Adenocarcinoma versus Adenosquamous
- ⇒ No: Other versus Adenosquamous.

*Controlling for ESTROGEN and SMOKING

Decision: retain or drop AGE from model.

We conclude that AGE is statistically significant for the Adenosquamous versus Adenocarcinoma comparison (category 1 vs. 0), but not for the Other versus Adenocarcinoma comparison (category 2 vs. 0), controlling for ESTROGEN and SMOKING.

The researcher must make a decision about whether to retain AGE 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.

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 $\alpha_g + \beta_{g1}X_1 + \beta_{g2}X_2 + \beta_{g3}X_1X_2$. The subscript g ($g = 1, 2$) indicates which comparison is being made (i.e., category 2 vs. 0, or category 1 vs. 0).

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 $g = 1, 2$

Likelihood ratio test

To test significance of interaction terms

$$H_0: \beta_{13} = \beta_{23} = 0$$

$$\text{Full model: } \alpha_g + \beta_{g1}X_1 + \beta_{g2}X_2 + \beta_{g3}X_1X_2$$

$$\text{Reduced model: } \alpha_g + \beta_{g1}X_1 + \beta_{g2}X_2$$

where $g = 1, 2$

Wald test

To test significance of interaction term at each level

$$H_0: \beta_{13} = 0$$

$$H_0: \beta_{23} = 0$$

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.

A full model with the interaction term would be fit and its likelihood compared against a reduced model without the interaction term.

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

Outcome variable has G levels:
 $(0, 1, 2, \dots, G-1)$

$$\ln \left[\frac{P(D=g | \mathbf{X})}{P(D=0 | \mathbf{X})} \right] = \alpha_g + \sum_{i=1}^p \beta_{gi} X_i$$

where $g = 1, 2, \dots, G-1$

Calculation of ORs and CIs as before

Likelihood ratio test	}	same procedures
Wald tests		

The model also easily extends for outcomes with more than three levels.

Assume that the outcome has G levels $(0, 1, 2, \dots, G-1)$. There are now $G-1$ possible comparisons with the reference category.

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 p independent variables times their β_g coefficients.

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.

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 $\text{df} = \text{number of parameters set to zero under } H_0 (= G-1 \text{ if } p = 1)$

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 $g = 1, 2, \dots, G-1$

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.)

Outcome with three levels

Consider probabilities of three outcomes:

$$P(D=0), P(D=1), P(D=2)$$

Logistic regression: dichotomous outcome

$$P(D=0|\mathbf{X}) = \frac{1}{1 + \exp\left[-(\alpha + \sum_{i=1}^p \beta_i X_i)\right]}$$

$$P(D=0|\mathbf{X}) = 1 - P(D=1|\mathbf{X})$$

Polytomous regression: three-level outcome

$$P(D=0|\mathbf{X}) + P(D=1|\mathbf{X}) + P(D=2|\mathbf{X}) = 1$$

$$h_1(\mathbf{X}) = \alpha_1 + \sum_{i=1}^p \beta_{1i} X_i$$

$$h_2(\mathbf{X}) = \alpha_2 + \sum_{i=1}^p \beta_{2i} X_i$$

$$\frac{P(D=1|\mathbf{X})}{P(D=0|\mathbf{X})} = \exp[h_1(\mathbf{X})]$$

$$\frac{P(D=2|\mathbf{X})}{P(D=0|\mathbf{X})} = \exp[h_2(\mathbf{X})]$$

We now present the likelihood function for polytomous logistic regression. This section may be omitted without loss of continuity.

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.

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 ($D=2$).

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.

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.

To simplify notation, we can let $h_1(\mathbf{X})$ be equal to α_1 plus the summation of the p independent variables times their β_1 coefficients and $h_2(\mathbf{X})$ be equal to α_2 plus the summation of the p independent variables times their β_2 coefficients.

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 $P(D=1|\mathbf{X})$ and $P(D=2|\mathbf{X})$ in terms of $P(D=0|\mathbf{X})$.

$$P(D=1|\mathbf{X}) = P(D=0|\mathbf{X}) \exp[h_1(\mathbf{X})]$$

$$P(D=2|\mathbf{X}) = P(D=0|\mathbf{X}) \exp[h_2(\mathbf{X})]$$

$$P(D=0|\mathbf{X}) + P(D=0|\mathbf{X}) \exp[h_1(\mathbf{X})] \\ + P(D=0|\mathbf{X}) \exp[h_2(\mathbf{X})] = 1$$

$$P(D=0|\mathbf{X})[1 + \exp h_1(\mathbf{X}) + \exp h_2(\mathbf{X})] = 1$$

With some algebra, we find that

$$P(D=0|\mathbf{X}) = \frac{1}{1 + \exp[h_1(\mathbf{X})] + \exp[h_2(\mathbf{X})]}$$

and that

$$P(D=1|\mathbf{X}) = \frac{\exp[h_1(\mathbf{X})]}{1 + \exp[h_1(\mathbf{X})] + \exp[h_2(\mathbf{X})]}$$

and that

$$P(D=2|\mathbf{X}) = \frac{\exp[h_2(\mathbf{X})]}{1 + \exp[h_1(\mathbf{X})] + \exp[h_2(\mathbf{X})]}$$

$L \Leftrightarrow$ joint probability of observed data.

The ML method chooses parameter estimates that maximize L

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.

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 $h_2(\mathbf{X})$.

These quantities can be substituted into the total probability equation and summed to 1.

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 $h_2(\mathbf{X})$.

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 $h_2(\mathbf{X})$.

The probability that the outcome is in category 2 can be found in a similar way, as shown on the left.

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: $j = 1, 2, 3, \dots, n$

$$y_{j0} = \begin{cases} 1 & \text{if outcome} = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$y_{j1} = \begin{cases} 1 & \text{if outcome} = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$y_{j2} = \begin{cases} 1 & \text{if outcome} = 2 \\ 0 & \text{otherwise} \end{cases}$$

$$P(D=0|\mathbf{X})^{y_{j0}} P(D=1|\mathbf{X})^{y_{j1}} P(D=2|\mathbf{X})^{y_{j2}}$$

$$y_{j0} + y_{j1} + y_{j2} = 1$$

since each subject has one outcome

$$\prod_{j=1}^n P(D=0|\mathbf{X})^{y_{j0}} P(D=1|\mathbf{X})^{y_{j1}} P(D=2|\mathbf{X})^{y_{j2}}$$

Likelihood for G outcome categories:

$$\prod_{j=1}^n \prod_{g=0}^{G-1} P(D=g|\mathbf{X})^{y_{jg}}$$

where $y_{jg} = \begin{cases} 1 & \text{if the } j\text{th subject has } D=g \\ & (g = 0, 1, \dots, G-1) \\ 0 & \text{if otherwise} \end{cases}$

Estimated α 's and β 's are those which maximize L

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.

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} .

Note that each individual subject contributes to only one of the category probabilities, since only one of the indicator variables will be nonzero.

The joint probability for the likelihood is the product of all the individual subject probabilities, assuming subject outcomes are independent.

The likelihood can be generalized to include G outcome categories by taking the product of each individual's contribution across the G outcome categories.

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 $0, 1, \dots, G-1$.

VII. Polytomous Versus Multiple Standard Logistic Regressions

Polytomous versus separate logistic models

Polytomous \Rightarrow uses data on all outcome categories in L .

Separate standard logistic \Rightarrow uses data on only two outcome categories at a time.



Parameter and variance estimates may differ.

Special case: One dichotomous predictor
Polytomous and standard logistic models
 \Rightarrow same estimates

One may wonder how using a polytomous model compares with using two or more separate dichotomous logistic models.

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.

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.

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 Chapter 10.

Detailed Outline

I. Overview (pages 270–271)

- A. Focus: modeling outcomes with more than two levels.
- B. Using previously described techniques by combining outcome categories.
- C. Nominal versus ordinal outcomes.

II. Polytomous logistic regression: An example with three categories (pages 272–275)

- 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):

$$\ln\left[\frac{P(D=1|X_1)}{P(D=0|X_1)}\right] = \alpha_1 + \beta_{11}X_1, \quad \ln\left[\frac{P(D=2|X_1)}{P(D=0|X_1)}\right] = \alpha_2 + \beta_{21}X_1.$$

III. Odds ratio with three categories (pages 275–279)

- 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^*) is

$$OR_g = \exp\left[(\beta_{g1}(X_1^{**} - X_1^*))\right] \text{ where } g = 1, 2.$$

IV. Statistical inference with three categories (pages 279–282)

- 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 for a 95% confidence interval for comparison of outcome level g versus 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 p predictors (pages 282–287)

- A. The model easily extends to include p independent variables.
- B. The general form of the model for G outcome levels is

$$\ln \left[\frac{P(D=g|X)}{P(D=0|X)} \right] = \alpha_g + \sum_{i=1}^p \beta_{gi} X_i \quad \text{where } g=1, 2, \dots, G-1.$$

- C. The calculation of the odds ratio, confidence intervals, and hypothesis testing using the likelihood ratio and Wald tests remain 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 288–290)

- A. For an outcome variable with G categories, the likelihood function is

$$\prod_{j=1}^n \prod_{g=0}^{G-1} P(D=g|X)^{y_{ig}} \quad \text{where } y_{ig} \begin{cases} 1 & \text{if the } j\text{th subject has } D=g \\ 0 & \text{otherwise} \end{cases}$$

where n is the total number of subjects and $g = 0, 1, \dots, G-1$

VII. Polytomous versus multiple standard logistic regressions (page 291)

- 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.

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 (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 \times 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) pertain 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 to a similar female. Consider the outcome comparison latent tuberculosis versus 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 to a similar female. Consider the outcome comparison active tuberculosis versus 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 to a similar female, with the outcome comparison active tuberculosis versus 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 versus 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 versus 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 versus 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 versus 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).

$$\ln \left[\frac{P(D=g|\mathbf{X})}{P(D=0|\mathbf{X})} \right] = \alpha_g + \beta_{g1}AGE + \beta_{g2}GENDER + \beta_{g3}SMOKE + \beta_{g4}HPT$$

where $g = 1, 2, 3$

Use the model to give an expression for the odds (severe versus 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 versus 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 versus 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

$$\ln \left[\frac{P(D = g | \mathbf{X})}{P(D = 0 | \mathbf{X})} \right] = \alpha_g + \beta_{g1} \text{VIRUS} + \beta_{g2} \text{AIDS} + \beta_{g3} \text{COMPLIANCE} + \beta_{g4} \text{AGE} + \beta_{g5} \text{GENDER} + \beta_{g6} \text{AIDSCOMP}$$

where $g = 1, 2$

2. Polytomous fitted model

$$\widehat{\ln} \left[\frac{P(D = 2 | \mathbf{X})}{P(D = 0 | \mathbf{X})} \right] = -2.82 + 1.35 \text{VIRUS} + 0.94 \text{AIDS} + 0.49 \text{COMPLIANCE} + 0.05 \text{AGE} + 0.41 \text{GENDER} + 0.33 \text{AIDSCOMP}$$

$$\widehat{\ln} \left[\frac{P(D = 1 | \mathbf{X})}{P(D = 0 | \mathbf{X})} \right] = -2.03 + 0.95 \text{VIRUS} + 0.76 \text{AIDS} + 0.34 \text{COMPLIANCE} + 0.03 \text{AGE} + 0.25 \text{GENDER} + 0.31 \text{AIDSCOMP}$$

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 Chapter 10).

4. $\widehat{OR}_{1 \text{ vs } 0} = \exp(0.25) = 1.28$

5. $\widehat{OR}_{2 \text{ vs } 0} = \exp(0.41) = 1.51$

6. $\widehat{OR}_{2 \text{ vs } 1} = \exp(0.41) / \exp(0.25) = \exp(0.16) = 1.17$

7. Two Wald statistics: $H_0: \beta_{16} = 0; z_1 = \frac{0.31}{0.17} = 1.82$; two-tailed P -value: 0.07

$H_0: \beta_{26} = 0; z_2 = \frac{0.33}{0.14} = 2.36$; two-tailed P -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):
for COMPLIANCE = 0: $\exp(0.94) = 2.56$
for COMPLIANCE = 1: $\exp(0.94 + 0.33) = 3.56$
9. $\widehat{OR} = \exp(1.35) = 3.86$; 95% CI: $\exp[1.35 \pm 1.96(0.11)] = (3.11, 4.79)$
10. Estimated odds = $\exp[-2.03 + (0.03)(20)] = \exp(-1.43) = 0.24$

10

Ordinal

Logistic

Regression

Contents	
Introduction	302
Abbreviated Outline	302
Objectives	303
Presentation	304
Detailed Outline	320
Practice Exercises	322
Test	324
Answers to Practice Exercises	325

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 304)
- II. Ordinal logistic regression: The proportional odds model (pages 304–310)
- III. Odds ratios and confidence limits (pages 310–313)
- IV. Extending the ordinal model (pages 314–316)
- V. Likelihood function for ordinal model (pages 316–317)
- VI. Ordinal versus multiple standard logistic regressions (pages 317–319)

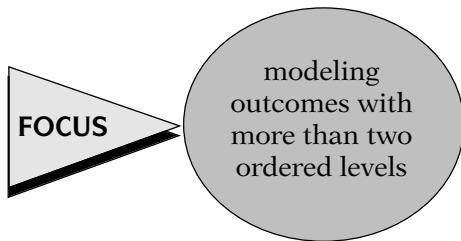
Objectives

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



Ordinal: levels have natural ordering

EXAMPLE

Tumor grade:

- well differentiated
- moderately differentiated
- poorly differentiated

Ordinal outcome \Rightarrow Polytomous model or
Ordinal model

Ordinal model takes into account order of
outcome levels

This presentation and the presentation in Chapter 9 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 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.

An ordinal outcome variable with three or more categories can be modeled with a polytomous model, as discussed in Chapter 9, but can also be modeled using ordinal logistic regression, provided that certain assumptions are met.

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

0	1	2	3	4
---	---	---	---	---

0	1	2	3	4
---	---	---	---	---

0	1	2	3	4
---	---	---	---	---

0	1	2	3	4
---	---	---	---	---

But, cannot allow

0	4	1	2	3
---	---	---	---	---

For G categories $\Rightarrow G-1$ ways to dichotomize outcome:

$$\begin{aligned} D \geq 1 &\text{ vs. } D < 1; \\ D \geq 2 &\text{ vs. } D < 2, \dots, \\ D \geq G-1 &\text{ vs. } D < G-1 \end{aligned}$$

$$\text{odds}(D \geq g) = \frac{P(D \geq g)}{P(D < g)}$$

where $g = 1, 2, 3, \dots, G-1$

Proportional odds assumption

EXAMPLE

$$\text{OR}(D \geq 1) = \text{OR}(D \geq 4)$$

Comparing two exposure groups
e.g., $E=1$ vs. $E=0$

where

$$\text{OR}_{(D \geq 1)} = \frac{\text{odds}[(D \geq 1) | E=1]}{\text{odds}[(D \geq 1) | E=0]}$$

$$\text{OR}_{(D \geq 4)} = \frac{\text{odds}[(D \geq 4) | E=1]}{\text{odds}[(D \geq 4) | E=0]}$$

Same odds ratio regardless of where categories are dichotomized

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.

More generally, if an ordinal outcome variable D has G categories ($D = 0, 1, 2, \dots, G-1$), then there are $G-1$ ways to dichotomize the outcome: ($D \geq 1$ vs. $D < 1$; $D \geq 2$ vs. $D < 2, \dots, D \geq G-1$ 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 ($g = 1, 2, 3, \dots, G-1$).

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.

In other words, the odds ratio is **invariant** to where the outcome categories are dichotomized.

Ordinal Variable	Parameter
------------------	-----------

Intercept	$\alpha_1, \alpha_2, \dots, \alpha_{G-1}$
X_1	β_1

Polytomous

Polytomous Variable	Parameter
Intercept	$\alpha_1, \alpha_2, \dots, \alpha_{G-1}$
X_1	$\beta_{11}, \beta_{21}, \dots, \beta_{(G-1)1}$

Odds are *not* invariant

EXAMPLE

$$\text{odds}(D \geq 1) \neq \text{odds}(D \geq 4)$$

where, for $E = 0$,

$$\text{odds}(D \geq 1) = \frac{P(D \geq 1 | E = 0)}{P(D < 1 | E = 0)}$$

$$\text{odds}(D \geq 4) = \frac{P(D \geq 4 | E = 0)}{P(D < 4 | E = 0)}$$

but

$$\text{OR}(D \geq 1) = \text{OR}(D \geq 4)$$

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.

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)

$$P(D \geq g | X_1) = \frac{1}{1 + \exp[-(\alpha_g + \beta_1 X_1)]}$$

where $g = 1, 2, \dots, G-1$

$$\begin{aligned} 1 - P(D \geq g | X_1) &= 1 - \frac{1}{1 + \exp[-(\alpha_g + \beta_1 X_1)]} \\ &= \frac{\exp[-(\alpha_g + \beta_1 X_1)]}{1 + \exp[-(\alpha_g + \beta_1 X_1)]} \\ &= P(D < g | X_1) \end{aligned}$$

We now present the form for the proportional odds model with an outcome (D) with G levels ($D = 0, 1, 2, \dots, G-1$) 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 $\beta_1 X_1$.

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

$$\text{odds} = \frac{P(D \geq g | X_1)}{1 - P(D \geq g | X_1)} = \frac{P(D \geq g | X_1)}{P(D < g | X_1)}$$

$$= \frac{\frac{1}{1 + \exp[-(\alpha_g + \beta_1 X_1)]}}{\frac{\exp[-(\alpha_g + \beta_1 X_1)]}{1 + \exp[-(\alpha_g + \beta_1 X_1)]}} = \exp(\alpha_g + \beta_1 X_1)$$

Proportional odds versus Standard logistic model:
 $P(D \geq g | \mathbf{X})$

The model can be defined equivalently in terms of the odds of an inequality. If we substitute the formula $P(D \geq g | X_1)$ 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 versus Polytomous model
model:

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 .

$$\begin{array}{ccc} \beta_1 & & \beta_{g1} \\ \nearrow \text{no } g \text{ subscript} & & \nearrow g \text{ subscript} \end{array}$$

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:

$$\text{odds} = \frac{P(D^* \leq g | X_1)}{P(D^* > g | X_1)} = \exp(\alpha_g^* - \beta_1^* X_1)$$

key differences

where $g = 1, 2, 3, \dots, G-1$
and $D^* = 1, 2, \dots, G$

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 $g = 1, 2, 3, \dots, G-1$ and where $D^* = 1, 2, \dots, G$. 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 $\alpha_g = -\alpha_g^*$.

Comparing formulations

$$\beta_1 = \beta_1^*$$

$$\text{but } \alpha_g = -\alpha_g^*$$

Formulation affects computer output

- SAS: consistent with first
- SPSS and Stata: consistent with alternative formulation

Advantage of ($D \geq g$):

consistent with formulations of standard logistic and polytomous models



For 2-level outcome ($D = 0, 1$), all three reduce to same model.

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).

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

A simple check of the proportional odds assumption:

	White	Black
Well + moderately differentiated	176	59
Poorly differentiated	31	22

$$\widehat{OR} = 2.12$$

	White	Black
Well differentiated	104	26
Moderately + poorly differentiated	103	55

$$\widehat{OR} = 2.14$$

Requirement: Collapsed ORs should be “close”

$E=0$ $E=1$

$D=0$	45	30
$D=1$	40	15
$D=2$	50	60

The 3×2 table of the data is presented on the left.

In order to examine the proportional odds assumption, the table is collapsed to form two other tables.

The first table combines the well-differentiated and moderately differentiated levels. The odds ratio is 2.12.

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.

Here is a different 3×2 table. This table will be collapsed in a similar fashion as the previous one.

	$E=0$	$E=1$		$E=0$	$E=1$
$D=0+1$	85	45	$D=0$	45	30
$D=2$	50	60	$D=1+2$	90	75
		$\widehat{OR} = 2.27$			$\widehat{OR} = 1.25$

Statistical test of assumption: **Score test**
Compares ordinal versus polytomous models

Test statistic $\sim \chi^2$ under H_0
with df = number of OR parameters tested

Alternate models for ordinal data

- continuation ratio
- partial proportional odds
- stereotype regression

The two collapsed table 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.

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.

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 $X_1 = 0$

$$\text{odds}(D \geq g) = \frac{P(D \geq g | X_1)}{P(D < g | X_1)} = \exp(\alpha_g + \beta_1 X_1)$$

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 $D \geq g$ versus $D < g$ 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. $X_1=0$).

$$\begin{aligned} \text{OR} &= \frac{\text{P}(D \geq g | X_1 = 1) / \text{P}(D < g | X_1 = 1)}{\text{P}(D \geq g | X_1 = 0) / \text{P}(D < g | X_1 = 0)} \\ &= \frac{\exp[\alpha_g + \beta_1(1)]}{\exp[\alpha_g + \beta_1(0)]} = \frac{\exp(\alpha_g + \beta_1)}{\exp(\alpha_g)} \\ &= e^{\beta_1} \end{aligned}$$

General case(levels X_1^{**} and X_1^* of X_1)

$$\begin{aligned} \text{OR} &= \frac{\exp(\alpha_g + \beta_1 X_1^{**})}{\exp(\alpha_g + \beta_1 X_1^*)} \\ &= \frac{\exp(\alpha_g) \exp(\beta_1 X_1^{**})}{\exp(\alpha_g) \exp(\beta_1 X_1^*)} \\ &= \exp[\beta_1(X_1^{**} - X_1^*)] \end{aligned}$$

CIs: same method as SLR to compute CIs**General case** (levels X_1^{**} and X_1^* of X_1)

$$95\% \text{ CI: } \exp\left[\hat{\beta}_i(X_1^{**} - X_1^*) \pm 1.96(X_1^{**} - X_1^*)s_{\hat{\beta}_i}\right]$$

EXAMPLE

Black/White Cancer Survival Study

Test of proportional odds assumption:

 H_0 : assumption holdsScore statistic: $\chi^2 = 0.0008$, df=1,
 $P = 0.9779$.

Conclusion: fail to reject null

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 $X_1=1$ versus $X_1=0$ in the comparison of disease levels $\geq g$ to levels $< g$ is then e to the β_1 .

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^* .

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.

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	-1.7388	0.1765
Intercept 2	-0.0089	0.1368
RACE	0.7555	0.2466

$$\widehat{OR} = \exp(0.7555) = 2.13$$

Interpretation of OR

Black versus white women with endometrial cancer over twice as likely to have more severe tumor grade:

$$\text{Since } \widehat{OR}(D \geq 2) = \widehat{OR}(D \geq 1) = 2.13$$

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.

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 versus moderately differentiated or well differentiated and over twice as likely as white women to have tumors classified as poorly differentiated or moderately differentiated versus 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;
 $g = 1, 2, 3, \dots, G-1$

$$\alpha_g > \alpha_{g+1}$$

↓

$$\alpha_1 > \alpha_2 > \dots > \alpha_{G-1}$$

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 $\alpha_1 > \alpha_2 > \dots > \alpha_{G-1}$.

Illustration

0	1	2	3	4
---	---	---	---	---

$$\alpha_1 = \log \text{odds } D \geq 1$$

0	1	2	3	4
---	---	---	---	---

$$\alpha_2 = \log \text{odds } D \geq 2$$

0	1	2	3	4
---	---	---	---	---

$$\alpha_3 = \log \text{odds } D \geq 3$$

0	1	2	3	4
---	---	---	---	---

$$\alpha_4 = \log \text{odds } D \geq 4$$

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**

$$\begin{aligned} 95\% \text{ CI} &= \exp[0.7555 \pm 1.96 (0.2466)] \\ &= (1.31, 3.45) \end{aligned}$$

Hypothesis testing

Likelihood ratio test or Wald test

$$H_0: \beta_1 = 0$$

Returning to our example, the 95% confidence interval for the OR for AGE is calculated as shown on the left.

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.

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.

Wald test

$$Z = \frac{0.7555}{0.2466} = 3.06, \quad P = 0.002$$

IV. Extending the Ordinal Model

$$P(D \geq g | \mathbf{X}) = \frac{1}{1 + \exp[-(\alpha_g + \sum_{i=1}^p \beta_i X_i)]}$$

where $g = 1, 2, 3, \dots, G-1$

Note: $P(D \geq 0 | \mathbf{X}) = 1$

$$\text{odds} = \frac{P(D \geq g | \mathbf{X})}{P(D < g | \mathbf{X})} = \exp(\alpha_g + \sum_{i=1}^p \beta_i X_i)$$

$\text{OR} = \exp(\beta_i)$, if X_i is coded (0, 1)

Expanding the model to add more independent variables is straightforward. The model with p independent variables is shown on the left.

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 p independent variable times its beta.

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.

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: $X_1 = \text{RACE}$ and $X_2 = \text{ESTROGEN}$.

EXAMPLE

$$D = \text{GRADE} = \begin{cases} 0 & \text{if well differentiated} \\ 1 & \text{if moderately differentiated} \\ 2 & \text{if poorly differentiated} \end{cases}$$

$$X_1 = \text{RACE} = \begin{cases} 0 & \text{if white} \\ 1 & \text{if black} \end{cases}$$

$$X_2 = \text{ESTROGEN} = \begin{cases} 0 & \text{if never user} \\ 1 & \text{if ever user} \end{cases}$$

EXAMPLE (continued)

$$P(D \geq g | \mathbf{X}) = \frac{1}{1 + \exp[-(\alpha_g + \beta_1 X_1 + \beta_2 X_2)]}$$

where $X_1 = \text{RACE} (0, 1)$
 $X_2 = \text{ESTROGEN} (0, 1)$
 $g = 1, 2$

$$\text{odds} = \frac{P(D \geq 2 | \mathbf{X})}{P(D < 2 | \mathbf{X})} = \exp(\alpha_2 + \beta_1 X_1 + \beta_2 X_2)$$

different α 's same β 's

$$\text{odds} = \frac{P(D \geq 1 | \mathbf{X})}{P(D < 1 | \mathbf{X})} = \exp(\alpha_1 + \beta_1 X_1 + \beta_2 X_2)$$

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

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 odds that the tumor grade is in a category greater than or equal to category 2 (i.e., poorly differentiated) versus 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) versus 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.

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.

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**

$$\widehat{OR} = \exp \hat{\beta}_1 = \exp(0.4270) = 1.53$$

95% confidence interval

$$\begin{aligned} 95\% \text{ CI} &= \exp[0.4270 \pm 1.96 (0.2720)] \\ &= (0.90, 2.61) \end{aligned}$$

Wald test

$$\begin{aligned} H_0: \beta_1 &= 0 \\ Z &= \frac{0.4270}{0.2720} = 1.57, \quad P = 0.12 \end{aligned}$$

Conclusion: fail to reject H_0

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.

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.

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 $P = 1/[1+(1/\text{odds})]$.

$$P(D=g) = [P(D \geq g)] - [P(D \geq g+1)]$$

For $g=2$

$$P(D=2) = P(D \geq 2) - P(D \geq 3)$$

Use relationship to obtain probability individual is in given outcome category.

L is product of individual contributions.

$$\prod_{j=1}^n \prod_{g=0}^{G-1} P(D=g | \mathbf{X})^{y_{ig}}$$

where

$$y_{ig} = \begin{cases} 1 & \text{if the } j\text{th subject has } D=g \\ 0 & \text{if otherwise} \end{cases}$$

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 an individual is in a specific outcome category for a given pattern of covariates (\mathbf{X}).

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 Versus 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.

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 versus less than 1, greater than or equal to 2 versus less than 2, and greater than or equal to 3 versus less than 3.

Original variable: 0, 1, 2, 3

Recoded:

≥ 1 vs. <1 , ≥ 2 vs. <2 , and ≥ 3 vs. <3

Three separate logistic regressions

Three sets of parameters

$$\alpha \geq 1 \text{ vs. } <1, \quad \beta \geq 1 \text{ vs. } <1$$

$$\alpha \geq 2 \text{ vs. } <2, \quad \beta \geq 2 \text{ vs. } <2$$

$$\alpha \geq 3 \text{ vs. } <3, \quad \beta \geq 3 \text{ vs. } <3$$

Logistic models Proportional odds model
(three parameters) (one parameter)

$$\beta \geq 1 \text{ vs. } <1$$

$$\beta \geq 2 \text{ vs. } <2$$

$$\beta \geq 3 \text{ vs. } <3$$

$$\beta$$

Is the proportional odds assumption met?

- Crude OR's "close"?
(No control of confounding)
- Beta coefficients in separate logistic models similar?
(Not a statistical test)

Is $\beta_{\geq 1 \text{ vs. } <1} \approx \beta_{\geq 2 \text{ vs. } <2} \approx \beta_{\geq 3 \text{ vs. } <3}$?

- Score test provides a test of proportional odds assumption

H_0 : assumption holds

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.

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.

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.

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.

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.

SUMMARY

✓ Chapter 10: 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.

Chapter 11: Logistic Regression for Correlated Data: GEE

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

All of the models presented thus far have assumed that observations are statistically independent, (i.e., are not correlated). In the next chapter (Chapter 11), we consider one approach for dealing with the situation in which study outcomes are not independent.

Detailed Outline

I. Overview (page 304)

- A. Focus: modeling outcomes with more than two levels.
- B. Ordinal outcome variables.

II. Ordinal logistic regression: The proportional odds model

(pages 304–310)

- 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 ($D = 0, 1, 2, \dots, G-1$) is

$$P(D \geq g | X_1) = \frac{1}{1 + \exp[-(\alpha_g + \beta_1 X_1)]} \quad \text{where } g = 1, 2, \dots, G-1$$

III. Odds ratios and confidence limits (pages 310–313)

- 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

$$\text{OR} = \exp[\beta_1(X_1^{**} - X_1^*)].$$
- 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

$$\exp[\hat{\beta}_1(X_1^{**} - X_1^*) \pm 1.96(X_1^{**} - X_1^*) s_{\hat{\beta}_1}].$$
- 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 314–316)

- A. The general form of the proportional odds model for G outcome categories and p independent variables is

$$1 - P(D \geq g | \mathbf{X}_1) = 1 - \frac{1}{1 + \exp[-(\alpha_g + \beta_1 X_1)]} \quad \text{where } g = 1, 2, \dots, G-1$$

- 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 316–317)

- 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

$$y_{jg} = \begin{cases} 1 & \text{if the } j\text{th subject has } D=g \\ 0 & \text{if otherwise} \end{cases}$$

where n is the total number of subjects, $g = 0, 1, \dots, G-1$ and $P(D = g|\mathbf{X}) = [P(D \geq g|\mathbf{X})] - [P(D \geq g+1|\mathbf{X})]$

VI. Ordinal versus multiple standard logistic regressions

(pages 317–319)

- 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.

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 (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 \times 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	-2.98	0.20
Intercept 2	-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 to 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 to 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 versus compliance. Consider the outcome comparison active tuberculosis versus 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 $P(D \geq 4)/P(D < 4)$ 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 $P(D \geq 3)/P(D < 3)$ is assumed equal to $P(D \geq 5)/P(D < 5)$.
- 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 $[P(D \geq 3|E=1)/P(D < 3|E=1)]/[P(D \geq 3|E=0)/P(D < 3|E=0)]$ is assumed equal to $[P(D \geq 5|E=1)/P(D < 5|E=1)]/[P(D \geq 5|E=0)/P(D < 5|E=0)]$.
- 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).

$$\ln \frac{P(D \geq g | \mathbf{X})}{P(D < g | \mathbf{X})} = \alpha_g + \beta_1 \text{AGE} + \beta_2 \text{GENDER} + \beta_3 \text{SMOKE} + \beta_4 \text{HPT}$$

where $g = 1, 2, 3$

Use the model to obtain an expression for the odds of a severe or critical outcome ($D \geq 2$) 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 versus 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 versus 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 versus nonhypertensive nonsmokers, controlling for AGE and GENDER.

Answers to Practice Exercises

1. Ordinal model

$$\ln \left[\frac{P(D \geq g | \mathbf{X})}{P(D < g | \mathbf{X})} \right] = \alpha_g + \beta_1 \text{VIRUS} + \beta_2 \text{AIDS} + \beta_3 \text{COMPLIANCE} + \beta_4 \text{AGE} + \beta_5 \text{GENDER} + \beta_6 \text{AIDSCOMP}$$

where $g = 1, 2$

2. Ordinal fitted model

$$\hat{\ln} \left[\frac{P(D \geq 2 | \mathbf{X})}{P(D < 2 | \mathbf{X})} \right] = -2.98 + 1.13 \text{VIRUS} + 0.82 \text{AIDS} + 0.38 \text{COMPLIANCE} + 0.04 \text{AGE} \\ + 0.35 \text{GENDER} + 0.31 \text{AIDSCOMP}$$

$$\hat{\ln} \left[\frac{P(D \geq 1 | \mathbf{X})}{P(D < 1 | \mathbf{X})} \right] = -1.65 + 1.13 \text{VIRUS} + 0.82 \text{AIDS} + 0.38 \text{COMPLIANCE} + 0.04 \text{AGE} \\ + 0.35 \text{GENDER} + 0.31 \text{AIDSCOMP}$$

3. $\widehat{OR} = \exp(0.35) = 1.42$
4. $\widehat{OR} = \exp(0.35) = 1.42$
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) versus compliant (COMPLIANCE=0) subjects:
 for AIDS = 0: $\exp(0.38) = 1.46$
 for AIDS = 1: $\exp(0.38 + 0.31) = 1.99$

11

Logistic Regression for Correlated Data: GEE

Contents	
Introduction	328
Abbreviated Outline	328
Objectives	329
Presentation	330
Detailed Outline	367
Practice Exercises	373
Test	374
Answers to Practice Exercises	375

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 is 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 (Chapter 12), examples are presented to illustrate the application and interpretation of GEE models. The final chapter in the text (Chapter 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.

- I. Overview (pages 330–331)
- II. An example (Infant Care Study) (pages 331–336)
- III. Data layout (page 337)
- IV. Covariance and correlation (pages 338–340)
- V. Generalized linear models (pages 341–344)
- VI. GEE models (pages 344–345)
- VII. Correlation structure (pages 345–348)
- VIII. Different types of correlation structure (pages 349–354)
- IX. Empirical and model-based variance estimators (pages 354–357)
- X. Statistical tests (pages 357–358)
- XI. Score equations and “score-like” equations (pages 359–361)
- XII. Generalizing the “score-like” equations to form GEE models (pages 362–366)

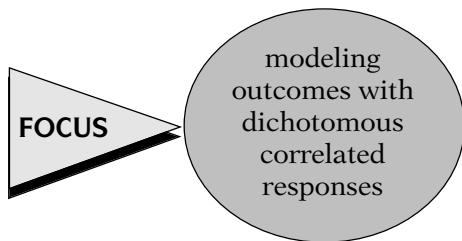
Objectives

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 household.
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:

Responses $\begin{cases} \text{correlated within clusters} \\ \text{independent between clusters} \end{cases}$

Ignoring within-cluster correlation



Incorrect inferences

A common assumption for correlated analyses is that the responses are correlated within the same cluster but are independent between different clusters.

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 versus standard logistic regression
(Ignores correlation)

- statistical inferences may differ
- similar use of output

Data source: Infant Care Study in Brazil

Subjects: 168 infants
136 with complete data

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.

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

$$D = \text{"Wasting"} \begin{cases} 1 & \text{if } z < -1 \\ 0 & \text{otherwise} \end{cases}$$

Independent variables:

BIRTHWGT (in grams)

GENDER

DIARRHEA $\begin{cases} 1 & \text{if symptoms present} \\ & \text{in past month} \\ 0 & \text{otherwise} \end{cases}$

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
<hr/>					
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
<hr/>					
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
<hr/>					

IDNO: identification number

MO: observation month

(provides order to subject-specific measurements)

OUTCOME: dichotomized z -score
(values can change month to month)

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

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.

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:

1. time-dependent
2. time-independent

Outcome variable generally varies within a cluster

Goal of analysis: to account for outcome variation within and between clusters

Model for Infant Care Study:

$$\text{logit } P(D=1 | \mathbf{X}) = \beta_0 + \beta_1 \text{BIRTHWGT} + \beta_2 \text{GENDER} + \beta_3 \text{DIARRHEA}$$

GEE Model

Variable	Coefficient	Empirical Std Err	Wald 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

- | | | |
|--------------------------------|---|-------------------|
| OR estimates | } | Use same formulas |
| Confidence intervals | | |
| Wald test statistics | | |
| Underlying assumptions | } | Differ |
| Method of parameter estimation | | |

Odds ratio

$$\widehat{\text{OR}}_{(\text{DIARRHEA} = 1 \text{ vs. DIARRHEA} = 0)} = \exp(0.2214) = 1.25$$

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. A correlated analysis attempts to account for the variation of the outcome from both within and between clusters.

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.

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.

The odds ratio comparing symptoms of diarrhea versus no diarrhea is calculated using the usual e to the $\hat{\beta}$ formula, yielding an estimated odds ratio of 1.25.

95% confidence interval

$$\begin{aligned}95\%CI &= \exp[0.2214 \pm 1.96(0.8558)] \\&= (0.23, 6.68)\end{aligned}$$

The 95% confidence interval is calculated using the usual large-sample formula, yielding a confidence interval of (0.23, 6.68).

Wald test

$$H_0: \beta_3 = 0$$

$$Z = \frac{0.2214}{0.8558} = 0.259, \quad P = 0.7958$$

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

Variable	Coefficient	Std Err	p-value	Wald
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 1203 separate observations across the 136 infants.

Odds ratio

$$\begin{aligned}\widehat{OR}_{(\text{DIARRHEA} = 1 \text{ vs. } \text{DIARRHEA} = 0)} \\= \exp(0.7764) = 2.17\end{aligned}$$

Using this output, the estimated odds ratio comparing symptoms of diarrhea versus no diarrhea is 2.17 for the naive model.

95% confidence interval

$$\begin{aligned}95\%CI &= \exp[0.7764 \pm 1.96(0.4538)] \\&= (0.89, 5.29)\end{aligned}$$

The 95% confidence interval for this odds ratio is calculated to be (0.89, 5.29).

Wald test

$$H_0: \beta_3 = 0$$

$$Z = \frac{0.7764}{0.4538} = 1.711, \quad P = 0.0871$$

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

2. P-Value of Wald test for BIRTHWGT

	GEE model	SLR model
P-Value	0.1081	0.0051

Why these differences?

GEE model: 136 independent clusters (infants)

Naive model: 1203 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

The Wald test statistic for DIARRHEA in the SLR model is calculated to be 1.711. The corresponding *P*-value is 0.0871.

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.

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.

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 1203 independent outcome measures.

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:



Framework for estimating

- correlations
- regression coefficients
- standard errors

To run a GEE analysis, the user specifies a correlation structure. The correlation structure provides a framework for the estimation of the correlation parameters, as well as estimation of the regression coefficients ($\beta_0, \beta_1, \beta_2, \dots, \beta_p$) 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 (section 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

S			0	Independent Variables				
u	R		u					
b	e		t					
j	p	T	c					
e	e	i	o					
c	a	m	m					
t	t	e	e					
(i)	(j)	(t_{ij})	Y_{ij}	X_{ij1}	X_{ij2}	...	X_{ijp}	
1	1	t_{11}	Y_{11}	X_{111}	X_{112}	...	X_{11p}	
1	2	t_{12}	Y_{12}	X_{121}	X_{122}	...	X_{12p}	
.	
.	
.	
1	n_1	t_{1n_1}	Y_{1n_1}	X_{1n_11}	X_{1n_12}	...	X_{1n_1p}	
...	
i	1	t_{i1}	Y_{i1}	X_{i11}	X_{i12}	...	X_{i1p}	
i	2	t_{i2}	Y_{i2}	X_{i21}	X_{i22}	...	X_{i2p}	
.	
.	
.	
i	n_i	t_{in_i}	Y_{in_i}	X_{in_i1}	X_{in_i2}	...	X_{in_ip}	
...	
K	1	t_{K1}	Y_{K1}	X_{K11}	X_{K12}	...	X_{K1p}	
K	2	t_{K2}	Y_{K2}	X_{K21}	X_{K22}	...	X_{K2p}	
.	
.	
.	
K	n_K	t_{Kn_K}	Y_{Kn_K}	X_{Kn_K1}	X_{Kn_K2}	...	X_{Kn_Kp}	

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, $X_{ij1}, X_{ij2}, \dots, X_{ijp}$.

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 which 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:

$$\widehat{\text{cov}}(X, Y) = \frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

Correlation

Population: $\rho_{xy} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}$

Sample: $r_{xy} = \frac{\widehat{\text{cov}}(X, Y)}{s_x s_y}$

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

$$X_1 = \text{height} \quad \text{cov}(X_2, Y) = 12 \text{ cov}(X_1, Y) \\ (\text{in feet})$$

$$X_2 = \text{height} \quad \text{BUT} \\ (\text{in inches})$$

$$Y = \text{weight} \quad \rho_{x_2y} = \rho_{x_1y}$$

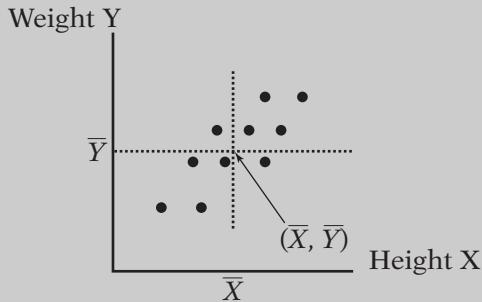
The correlation is a standardized measure of covariance in which the units of X and Y are the standard deviations of X and Y , respectively. The actual units used for the value of variables affect measures of covariance but not measures of correlation, which are scale-free. For example, the covariance 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

Height and weight



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 are 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.

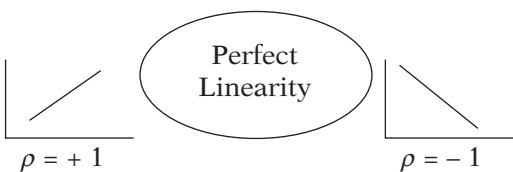
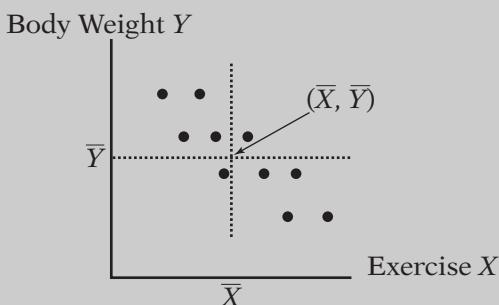
Negative correlation

On average, as X gets larger, Y gets smaller; or as X gets smaller, Y gets larger.

EXAMPLE

Hours of exercise and body weight

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

$$Y = \beta_0 + \beta_1 X, \text{ for a given } X$$

X and Y independent $\Rightarrow \rho = 0$

BUT

$$\rho = 0 \Rightarrow \begin{cases} X \text{ and } Y \text{ independent} \\ \text{or} \\ X \text{ and } Y \text{ have nonlinear relationship} \end{cases}$$

Correlations on same variable

$$(Y_1, Y_2, \dots, Y_n)$$

$$\rho_{Y_1 Y_2}, \rho_{Y_1 Y_3}, \dots, \text{etc.}$$

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 .

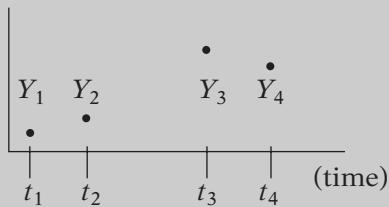
We have been discussing correlation in terms of two different variables such as height and weight. We can also consider correlations between repeated observations (Y_1, Y_2, \dots, Y_n) on the same variable Y .

EXAMPLE

Systolic blood pressure on same individual over time

Expect $\rho_{Y_j Y_k} >$ for some j, k

Also,



Expect $\rho_{Y_1 Y_2}$ or $\rho_{Y_3 Y_4} >$

$$\rho_{Y_1 Y_3}, \rho_{Y_1 Y_4}, \rho_{Y_2 Y_3}, \rho_{Y_2 Y_4},$$

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.

EXAMPLE

Daily inhaler use (1=yes, 0=no) on same individual over time

Expect $\rho_{Y_j Y_k} > 0$ for same subject

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).

V. Generalized Linear Models

General form of many statistical models:

$$Y = f(X_1, X_2, \dots, X_p) + \epsilon$$

where: Y is random

X_1, X_2, \dots, X_p are fixed
 ϵ is random

Specify:

1. a function (f) for the fixed predictors
2. a distribution for the random error (ϵ)

GLM models include:

logistic regression
linear regression
Poisson regression

GEE models are extensions of GLM

GLM: a generalization of the classical linear model

Linear regression

Outcome

- continuous
- normal distribution

Logistic regression

Outcome

- dichotomous
- binomial distribution:
 $E(Y) = \mu = P(Y=1)$

Logistic regression used to model

$$P(Y|X_1, X_2, \dots, X_p)$$

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 $Y = f(X_1, X_2, \dots, X_p) + \epsilon$, where Y is the response variable, X_1, X_2, \dots, X_p are the predictor variables, and ϵ represents random error. In this framework, the model for Y consists of a fixed component [$f(X_1, X_2, \dots, X_p)$] and a random component (ϵ). A function (f) for the fixed predictors and a distribution for the random error (ϵ) are specified.

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 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.

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., $P(Y=1)$]. It is this probability that is modeled in logistic regression.

Exponential family distributions include:

- binomial
- normal
- Poisson
- exponential
- gamma

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

Three components for GLM

1. Random component
2. Systematic component
3. Link function

1. Random component

Y follows a distribution from the exponential family

2. Systematic component

The X 's are combined in the model linearly, (i.e., $\beta_0 + \sum \beta_h X_h$)

Logistic model:

$$P(X) = \frac{1}{1 + \exp[-(\beta_0 + \sum \beta_h X_h)]}$$

linear component



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.

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.

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. The **random component** is that the outcome (Y) is required 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. The **systematic component** requires that the X 's be combined in the model as a linear function ($\beta_0 + \sum \beta_h X_h$) 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 $\beta_0 + \sum \beta_h X_h$

Logistic regression (logit link)

$$g(\mu) = \log\left[\frac{\mu}{1-\mu}\right] = \text{logit}(\mu)$$

Alternate formulation

Inverse of link function = g^{-1} satisfies

$$g^{-1}(g(\mu)) = \mu$$

Inverse of logit function in terms of $(\mathbf{X}, \boldsymbol{\beta})$

$$g^{-1}(\mathbf{X}, \boldsymbol{\beta}) = \mu = \frac{1}{1 + \exp\left[-(\alpha + \sum_{h=1}^p \beta_h X_h)\right]}$$

where

$$g(\mu) = \text{logit } P(D=1 | \mathbf{X}) = \beta_0 + \sum_{h=1}^p \beta_h X_h$$

GLM

- uses ML estimation
- requires likelihood function L where

$$L = \prod_{i=1}^h L_i$$

(assumes Y_i are independent)

If Y_i not independent and not normal

↓

L complicated or intractable

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.

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].

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 ($\boldsymbol{\beta}$) (i.e., $g^{-1}(\mathbf{X}, \boldsymbol{\beta})$). 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 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. However, if the responses are not independent, then the likelihood can become complicated, or intractable.

If Y_i not independent but MV normal



L specified

If Y_i not independent and *not* MV normal



Quasi-likelihood theory

Quasi-likelihood

- no likelihood
- specify mean variance relationship
- foundation of GEE

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 to model development.

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

$$g(\mu) = \beta_0 + \sum_{h=1}^p \beta_h X_h$$

For $Y(0, 1) \Rightarrow$ logit link

$$g(\mu) = \text{logit } P(Y=1 | \mathbf{X}) = \beta_0 + \sum_{h=1}^p \beta_h X_h$$

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 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 $P(Y=1 | \mathbf{X})$ equals β_0 plus the summation of the p independent variables times their beta coefficients.

Correlated versus Independent

- identical model
but
- different assumptions

GEE

- generalization of quasi-likelihood
- specify a “working” correlation structure for within-cluster correlations
- assume independence between clusters

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 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.

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.

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

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

$$\mathbf{V} = \begin{bmatrix} \text{var}(Y_1) & \text{cov}(Y_1, Y_2) \\ \text{cov}(Y_1, Y_2) & \text{var}(Y_2) \end{bmatrix}$$

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

$$\mathbf{C} = \begin{bmatrix} 1 & \text{corr}(Y_1, Y_2) \\ \text{corr}(Y_1, Y_2) & 1 \end{bmatrix}$$

Diagonal matrix: has 0 in all nondiagonal entries.

Diagonal 2×2 matrix with variances on diagonal

$$\mathbf{D} = \begin{bmatrix} \text{var}(Y_1) & 0 \\ 0 & \text{var}(Y_2) \end{bmatrix}$$

Can extend to $N \times N$ matrices

Matrices symmetric: $(i, j) = (j, i)$ element

$$\begin{aligned} \text{cov}(Y_1, Y_2) &= \text{cov}(Y_2, Y_1) \\ \text{corr}(Y_1, Y_2) &= \text{corr}(Y_2, Y_1) \end{aligned}$$

Relationship between covariance and correlation expressed as

$$\begin{aligned} \text{cov}(Y_1, Y_2) &= \sqrt{\text{var}(Y_1)}[\text{corr}(Y_1, Y_2)]\sqrt{\text{var}(Y_2)} \end{aligned}$$

$$\text{Matrix version: } \mathbf{V} = \mathbf{D}^{\frac{1}{2}} \mathbf{C} \mathbf{D}^{\frac{1}{2}}$$

$$\text{where } \mathbf{D}^{\frac{1}{2}} \times \mathbf{D}^{\frac{1}{2}} = \mathbf{D}$$

Logistic regression

$$\mathbf{D} = \begin{bmatrix} \mu_1(1-\mu_1) & 0 \\ 0 & \mu_2(1-\mu_2) \end{bmatrix}$$

where

$$\text{var}(Y_i) = \mu_i(1-\mu_i)$$

$$\mu_i = g^{-1}(\mathbf{X}, \beta)$$

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.

A **diagonal matrix** has a 0 in all nondiagonal entries.

A 2×2 diagonal matrix (**D**) with the variances along the diagonal is of the form shown at left.

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.

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 .

The relationship between covariance and correlation can be similarly be expressed in terms of the matrices **V**, **C**, and **D** as shown on the left.

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 j th

and k th response from subject $i = \rho_{ijk}$

Between-subject correlations = 0

1	ρ_{12}	ρ_{13}	ρ_{14}	0	0	0	0	0	0	0	0
ρ_{12}	1	ρ_{13}	ρ_{14}	0	0	0	0	0	0	0	0
ρ_{13}	ρ_{12}	1	ρ_{14}	0	0	0	0	0	0	0	0
ρ_{14}	ρ_{12}	ρ_{13}	1	0	0	0	0	0	0	0	0
	0	0	0	0	1	ρ_{212}	ρ_{213}	ρ_{214}	0	0	0
	0	0	0	0	ρ_{212}	1	ρ_{223}	ρ_{224}	0	0	0
	0	0	0	0	ρ_{213}	ρ_{223}	1	ρ_{234}	0	0	0
	0	0	0	0	ρ_{214}	ρ_{224}	ρ_{234}	1	0	0	0
	0	0	0	0	0	0	0	0	1	ρ_{312}	ρ_{313}
	0	0	0	0	0	0	0	0	ρ_{312}	1	ρ_{323}
	0	0	0	0	0	0	0	0	ρ_{323}	ρ_{324}	1
	0	0	0	0	0	0	0	0	ρ_{314}	ρ_{324}	ρ_{334}
	0	0	0	0	0	0	0	0	0	0	1

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 j th and k th response from subject i) is ρ_{ijk} .

Block diagonal matrix: subject-specific correlation matrices form blocks (B_i)

$$\begin{bmatrix} B_1 & \mathbf{0} \\ \mathbf{B}_2 & \\ \mathbf{0} & B_3 \end{bmatrix} \text{ where } B_i = i\text{th block}$$

This correlation matrix is called a **block diagonal matrix**, where subject-specific correlation matrices are the blocks along the diagonal of the matrix.

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.

EXAMPLE

18 ρ 's (6 per cluster/subject) but 12 observations

Subject i : $\{\rho_{i12}, \rho_{i13}, \rho_{i14}, \rho_{i23}, \rho_{i24}, \rho_{i34}\}$

parameters > # observations $\Rightarrow \hat{\beta}_i$ not valid

GEE approach: common set of ρ 's 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 over-parameterized 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

1	ρ_{12}	ρ_{13}	ρ_{14}	0	0	0	0	0	0	0	0
ρ_{12}	1	ρ_{23}	ρ_{24}	0	0	0	0	0	0	0	0
ρ_{13}	ρ_{23}	1	ρ_{34}	0	0	0	0	0	0	0	0
ρ_{14}	ρ_{24}	ρ_{34}	1	0	0	0	0	0	0	0	0
0	0	0	0	1	ρ_{12}	ρ_{13}	ρ_{14}	0	0	0	0
0	0	0	0	ρ_{12}	1	ρ_{23}	ρ_{24}	0	0	0	0
0	0	0	0	ρ_{13}	ρ_{23}	1	ρ_{34}	0	0	0	0
0	0	0	0	ρ_{14}	ρ_{24}	ρ_{34}	1	0	0	0	0
0	0	0	0	0	0	0	0	1	ρ_{12}	ρ_{13}	ρ_{14}
0	0	0	0	0	0	0	0	ρ_{12}	1	ρ_{23}	ρ_{24}
0	0	0	0	0	0	0	0	ρ_{13}	ρ_{23}	1	ρ_{34}
0	0	0	0	0	0	0	0	ρ_{14}	ρ_{24}	ρ_{34}	1

Now only 6 ρ 's for 12 observations:

\downarrow # ρ 's by factor of 3 (= # subjects)

In general, for K subjects:

$\rho_{ijk} \Rightarrow \rho_{jk}$: # of ρ 's \downarrow by factor of K

Example above: **unstructured** correlation structure

Next section shows other structures.

Using the previous example, 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, a common set of correlation parameters for K subjects reduces the number of correlation parameters by a factor of K .

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

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Exchangeable

Assumption: any two responses within a cluster have same correlation (ρ)

With five responses per cluster

$$\begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix}$$

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 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 parameter(s) are used 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.

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.

$K = 14$ schools

n_i = # students from school i

$$\sum_{i=1}^k n_i = 237$$

School i : exchange order #2 \leftrightarrow # 9

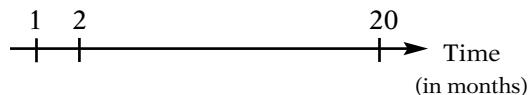
Will not affect analysis

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 $k = 1, 2, \dots, m$
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 ρ 's

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_i responses $\Rightarrow n(n-1)/2$ distinct ρ 's

$$\rho_{jk} \neq \rho_{j'k'} \text{ unless } j = j' \text{ and } k = k'$$

$$\rho_{12} \neq \rho_{34} \text{ even if } t_2 - t_1 = t_4 - t_3$$

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 $k = 1, 2, \dots, m$, 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.

Unlike the correlation structures discussed previously, 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 $i \neq i'$

$$\rho_{A12} = \rho_{B12} = \rho_{12}$$



different clusters

Order $\{Y_{i1}, Y_{i2}, \dots, Y_{ik}\}$ not arbitrary
(e.g., cannot switch Y_{A1} and Y_{A4}
unless all Y_{i1} and Y_{i4} switched).

Fixed

User specifies fixed values for ρ .

$\rho = 0.1$ for first and fourth responses;
0.3 otherwise

$$\begin{bmatrix} 1.0 & 0.3 & 0.3 & 0.1 \\ 0.3 & 1.0 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1.0 & 0.3 \\ 0.1 & 0.3 & 0.3 & 1.0 \end{bmatrix}$$

No ρ estimated.

Choice of structure not always clear.

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 correlation structure.

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.

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.

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.

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.

IX. Empirical and Model-Based Variance Estimators

GEE estimates have desirable *asymptotic* properties.



$K \rightarrow \infty$ (i.e., K “large”)

where $K = \# \text{ clusters}$

“Large” is subjective

Two statistical properties of GEE estimates (if model correct):

1. **Consistent**
 $\hat{\beta} \rightarrow \beta$ as $K \rightarrow \infty$
2. **Asymptotically normal**
 $\hat{\beta} \sim \text{normal}$ as $K \rightarrow \infty$

Asymptotic normal property allows:

- confidence intervals
- statistical tests

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.

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.

To correctly specify a GEE model:

- specify correct $g(\mu)$
- specify correct \mathbf{C}_i

$\hat{\beta}_h$ consistent even if \mathbf{C}_i misspecified
but

$\hat{\beta}_h$ more efficient if \mathbf{C}_i correct

To construct CIs, need $\widehat{\text{var}}(\hat{\beta})$

Two types of variance estimators:

- model-based
- empirical

No effect on $\hat{\beta}$

Effect on $\widehat{\text{var}}(\hat{\beta})$

Model-based variance estimators:

- similar in form to variance estimators in GLM
- consistent only if \mathbf{C}_i correctly specified

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).

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.

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 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 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 versus estimation of $\text{var}(\hat{\beta})$

- 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 $\text{var}(\hat{\beta})$ 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 Section 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 Section 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

Under H_0 , test statistics approximate χ^2 with df = number of parameters tested.

To test one $\hat{\beta}_h$, the Wald test statistic is of the familiar form

$$Z = \frac{\hat{\beta}_h}{s_{\hat{\beta}_h}}$$

Next two sections:

- GEE theory
- use calculus and matrix notation

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.

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 $\hat{\beta}_h$.

The use of the Score test, Wald test, and generalized Wald test will be further illustrated in the examples presented in the Chapter 12.

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**.

$$\left. \begin{array}{l} S_1 = \frac{\partial \ln L}{\partial \beta_0} = 0 \\ S_2 = \frac{\partial \ln L}{\partial \beta_1} = 0 \\ \vdots \\ S_{p+1} = \frac{\partial \ln L}{\partial \beta_p} = 0 \end{array} \right\} \begin{array}{l} p+1 \text{ equations in} \\ p+1 \text{ unknowns } (\beta's) \end{array}$$

In GLM, score equations involve $\mu_i = E(Y_i)$ and $\text{var}(Y_i)$

K = # of subjects

$p+1$ = # of parameters ($\beta_h, h=0,1,2, \dots, p$)

Yields $p+1$ score equations (S_1, S_2, \dots, S_{p+1}):

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.

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 ($\beta_0, \beta_1, \beta_2, \dots, \beta_p$), and the observed data.

If there are K subjects, with each subject contributing one response, and $p+1$ beta parameters ($\beta_0, \beta_1, \beta_2, \dots, \beta_p$), 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.

$$S_{h+1} = \sum_{i=1}^K \frac{\partial \mu_i}{\beta_h} [\text{var}(Y_i)]^{-1} [Y_i - \mu_i] = 0$$

↑ ↑ ↑
partial derivative variance residual

Solution: iterative (by computer)

GLM score equations:

- completely specified by $E(Y_i)$ and $\text{var}(Y_i)$
- basis of QL estimation

QL estimation:

- **“score-like” equations**
- No likelihood
- $\text{var}(Y_i) = \phi V(\mu_i)$
 - ↑ ↑
scale function of μ
- $g(\mu) = \beta_0 + \sum_{h=1}^p \beta_h X_h$
- solution yields QL estimates

Logistic regression: $Y = (0, 1)$

$$\mu = P(Y=1|\mathbf{X})$$

$$\begin{aligned} V(\mu) &= P(Y=1|\mathbf{X})[1-P(Y=1|\mathbf{X})] \\ &= \mu(1-\mu) \end{aligned}$$

The $(h+1)$ st score equation (S_{h+1}) is written as shown on the left. For each score equation, the i th 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.

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.

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.

For logistic regression, in which the outcome is coded 0 or 1, the mean response is the probability of obtaining the event, $P(Y=1|\mathbf{X})$. The variance of the response equals $P(Y=1|\mathbf{X})$ times 1 minus $P(Y=1|\mathbf{X})$. 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

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.

Equations	Allow extra variation?
QL: “score-like”	Yes
GLM: score	No

Summary: ML Versus 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 $\frac{\partial \ln L}{\partial \beta}$	—
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:

β and α
 ↗ ↘
 regression correlation
 parameters parameters

Matrix notation used to describe GEE

Matrices needed specific to each subject (cluster): \mathbf{Y}_i , $\boldsymbol{\mu}_i$, \mathbf{D}_i , \mathbf{C}_i , and \mathbf{W}_i

$$\mathbf{Y}_i = \begin{Bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{Bmatrix} \quad \text{vector of } i\text{th subject's observed responses}$$

$$\boldsymbol{\mu}_i = \begin{Bmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{in_i} \end{Bmatrix} \quad \text{vector of } i\text{th subject's mean responses}$$

\mathbf{C}_i = working correlation matrix ($n_i \times n_i$)

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.

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 and vectors are indicated by the use of bold letters. The matrices that are needed are specific for each subject (i.e., i th 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:

\mathbf{Y}_i is the vector (i.e., collection) of the i th subject's observed responses.

$\boldsymbol{\mu}_i$ is a vector of the i th 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 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{W}_i = working covariance matrix ($n_i \times n_i$)

$$\mathbf{W}_i = \phi \mathbf{D}_i^{\frac{1}{2}} \mathbf{C}_i \mathbf{D}_i^{\frac{1}{2}}$$

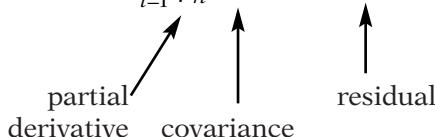
GEE: form similar to score equations

If K = # of subjects

n_i = # responses of subject i

$p+1$ = # of parameters (β_h ; $h=0, 1, 2, \dots, p$)

$$\text{GEE}_{h+1}: \sum_{i=1}^K \frac{\partial \boldsymbol{\mu}_i}{\partial \beta_h} [\mathbf{W}_i]^{-1} [\mathbf{Y}_i - \boldsymbol{\mu}_i] = 0$$



where

$$\mathbf{W}_i = \phi \mathbf{D}_i^{\frac{1}{2}} \mathbf{C}_i \mathbf{D}_i^{\frac{1}{2}}$$

Yields $p+1$ GEE equations of the above form

\mathbf{D}_i is a diagonal matrix whose j th diagonal (representing the j th observation of the i th 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 is an $n_i \times n_i$ variance-covariance matrix for the i th 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 .

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 ($\beta_0, \beta_1, \beta_2, \dots, \beta_p$), 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 i th 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 versus 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 i th subject's responses.

GEE model parameters—three types:

1. Regression parameters (β)

Express relationship between predictors and outcome.

There are three types of parameters in a GEE model. These are as follows.

2. Correlation parameters (α)

Express within-cluster correlation; user specifies \mathbf{C}_i .

1. The **regression parameters** (β) 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. The **correlation parameters** (α) 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

α and β estimated iteratively:

Estimates updated alternately

\Rightarrow convergence

To run GEE model, specify:

- $g(\mu)$ = link function
- $V(\mu)$ = mean variance relationship
- C_i = working correlation structure

GLM—no specification of a 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

β_i interpretation same as SLR

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).

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.

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.

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 Versus 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.

SUMMARY

- ✓ Chapter 11: 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 12: GEE Examples

In the next chapter (Chapter 12), 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 330–331)

- 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 331–336)

- 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 337)

- A. For repeated measures for K subjects:
 - i. the i th subject has n_i measurements recorded;
 - ii. 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, $X_{ij1}, X_{ij2}, \dots, X_{ijp}$.
- 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 338–340)

- 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.

$$\rho_{xy} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

- i. correlation values range from -1 to $+1$;
- ii. can have correlations between observations on the same 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 341–344)

- 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 $[g^{-1}(\mathbf{X}, \boldsymbol{\beta})] = \mu$;
 - c. for logistic regression, the inverse of the logit function is the familiar logistic model for the probability of an event:

$$g^{-1}(\mathbf{X}, \boldsymbol{\beta}) = \mu = \frac{1}{1 + \exp\left[-(\alpha + \sum_{i=1}^p \beta_i X_i)\right]}$$

- 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 344–345)

- 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:
$$g(\mu) = \text{logit}[\mathbf{P}(D=1 | \mathbf{X})] = \beta_0 + \sum_{i=1}^p \beta_i X_i$$
- 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 345–348)

- 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 349–354)

- 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 $k = 1, 2, \dots, m$, 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 354–357)

- 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 357–358)**A. Score test**

- i. the test statistic is based on the “score-like” equations;
- ii. under the null, the test statistic is distributed approximate 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 approximate 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 359–361)

- 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

$$\sum_{i=1}^K \frac{\partial \mu_i}{\partial \beta_h} [\text{var}(Y_i)]^{-1} [Y_i - \mu_i] = 0$$

where $h = (0, 1, 2, \dots, p)$;

- 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 are 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 is not formulated.
 - ii. for a dichotomous outcome with a binomial distribution, $\text{var}(Y) = \phi V(\mu)$, 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 362–366)

- 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:

$$\sum_{i=1}^K \frac{\partial \boldsymbol{\mu}'_i}{\beta_h} [\mathbf{W}_i]^{-1} [\mathbf{Y}_i - \boldsymbol{\mu}_i] = 0$$

where $\mathbf{W}_i = \phi \mathbf{D}_i^{\frac{1}{2}} \mathbf{C}_i \mathbf{D}_i^{\frac{1}{2}}$ and where $h = 0, 1, 2, \dots, p$.

- 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 (β): these express the relationship between the predictors and the outcome. In logistic regression, the betas allow estimation of odds ratios.
- ii. correlation parameters (α): 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.

Practice Exercises

Questions 1–5 pertain to identifying the following correlation structures that apply to clusters of four responses each:

$$\text{A} \quad \begin{bmatrix} 1 & 0.27 & 0.27 & 0.27 \\ 0.27 & 1 & 0.27 & 0.27 \\ 0.27 & 0.27 & 1 & 0.27 \\ 0.27 & 0.27 & 0.27 & 1 \end{bmatrix} \quad \text{B} \quad \begin{bmatrix} 1 & 0.35 & 0 & 0 \\ 0.35 & 1 & 0.35 & 0 \\ 0 & 0.35 & 1 & 0.35 \\ 0 & 0 & 0.35 & 1 \end{bmatrix} \quad \text{C} \quad \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\text{D} \quad \begin{bmatrix} 1 & 0.50 & 0.25 & 0.125 \\ 0.50 & 1 & 0.50 & 0.25 \\ 0.25 & 0.50 & 1 & 0.50 \\ 0.125 & 0.25 & 0.50 & 1 \end{bmatrix} \quad \text{E} \quad \begin{bmatrix} 1 & 0.50 & 0.25 & 0.125 \\ 0.50 & 1 & 0.31 & 0.46 \\ 0.25 & 0.31 & 1 & 0.163 \\ 0.125 & 0.46 & 0.163 & 1 \end{bmatrix}$$

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 $\text{var}(\hat{\beta})$] 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 non-linear 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.

12

GEE

Examples

Contents	
Introduction	378
Abbreviated Outline	378
Objectives	379
Presentation	380
Detailed Outline	396
Practice Exercises	397
Test	400
Answers to Practice Exercises	402

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.

- I. Overview (page 380)
- II. Example I: Infant Care Study (pages 380–388)
- III. Example II: Aspirin–Heart Bypass Study (pages 389–393)
- IV. Example III: Heartburn Relief Study (pages 393–395)

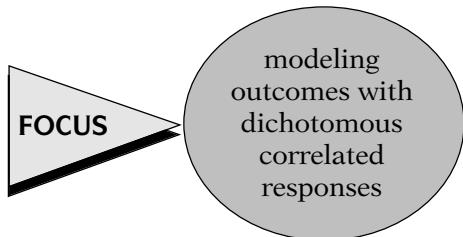
Objectives

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 to a standard logistic regression model.

Presentation

I. Overview



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 Chapter 11

In Chapter 11, 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

$$D = \text{"Wasting"} \quad \begin{cases} 1 & \text{if } z < -1 \\ 0 & \text{otherwise} \end{cases}$$

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

$$\text{DIARRHEA} = \begin{cases} 1 & \text{if symptoms present} \\ & \text{in past month} \\ 0 & \text{otherwise} \end{cases}$$

DIARRHEA

- exposure of interest
- time-dependent variable

Infant Care Study Model

$$\text{logit } P(D=1|\mathbf{X}) = \beta_0 + \beta_1 \text{BIRTHWGT} + \beta_2 \text{GENDER} + \beta_3 \text{DIARRHEA}$$

Five GEE models presented, with different \mathbf{C}_i :

1. AR1 autoregressive
2. Exchangeable
3. Fixed
4. Independent
5. Independent (SLR)

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.

The model for the study can be stated as shown on the left.

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 8.2.

Output presented:

- $\hat{\beta}_h$, $S_{\hat{\beta}}$ (empirical), and Wald test P-values
- “working” correlation matrix (\mathbf{C}_i) containing $\hat{\rho}$

Sample:

$K = 168$ infants, $n_i \leq 9$, but

9 infants “exposed cases”:
 (i.e., $D = 1$ and DIARRHEA = 1
 for any month)

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 Chapter 11).

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.

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

Variable	Coefficient	Empirical Std Err	Wald p-value
INTERCEPT	-1.3978	1.1960	0.2425
BIRTHWT	-0.0005	0.0003	0.1080
GENDER	0.0024	0.5546	0.9965
DIARRHEA	0.2214	0.8558	0.7958

Effect of DIARRHEA:

$$\widehat{OR} = \exp(0.2214) = 1.25$$

$$95\% \text{ CI} = \exp[0.2214 \pm 1.96(0.8558)] \\ = (0.23, 6.68)$$

Working correlation matrix: 9×9

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.

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 $\exp[0.2214 \pm 1.96(0.8558)]$, yielding a confidence interval of (0.23, 6.68).

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.

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.

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

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

$$\hat{\beta}_3 \text{ for DIARRHEA} = 0.6485$$

(versus 0.2214 with 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 $\hat{\rho}$: $\hat{\rho} = 0.4381$

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 to the parameter estimate of 0.2214 for the same coefficient using the AR1 correlation structure in Model 1.

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

Correlation structure (fixed) for Model 3:
combines AR1 and exchangeable features

Choice of ρ at discretion of user,
but may not always converge

Allows flexibility specifying complicated \mathbf{C}_i

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.

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.

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.

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.

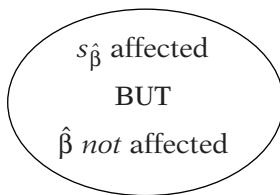
Independent correlation structure: two models

Model 4: uses empirical $s_{\hat{\beta}}$;

ϕ not fixed

Model 5: uses model-based $s_{\hat{\beta}}$;

ϕ fixed at 1



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.

Model 4: independent correlation structure

Variable	Coefficient	Empirical Std Err	Wald p-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

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.

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.

The outputs for Model 4 and Model 5 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
BIRTHWT	-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.

Model 4 vs. Model 5

- Parameter estimates same
- $s_{\hat{\beta}}$ Model 4 > $s_{\hat{\beta}}$ Model 5

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 to 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.

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

$$\begin{array}{l} \text{(usually)} \longrightarrow s_{\hat{\beta}} \\ \longrightarrow \widehat{OR} \end{array}$$

For Models 1–5:

$$\widehat{OR} \text{ range} = 1.25–3.39$$

\widehat{OR} range suggests model instability.

Instability likely due to small number (nine) of exposed cases.

Which models to eliminate?

Models 4 and 5 (independent):
evidence of correlated observations

Model 2 (exchangeable):
if autocorrelation suspected

Remaining models: similar results:

Model 1 (AR1)
 \widehat{OR} (95% CI) = 1.25 (0.23, 6.68)

Model 3 (fixed)
 \widehat{OR} (95% CI) = 1.29 (0.26, 6.46)

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).

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.

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.

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.

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)

Correlation structures to consider:

- exchangeable
- independent

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).

Model 1: interaction model

Interaction terms between ASPIRIN and the other four covariates included.

$$\begin{aligned} \text{logit } P(D=1|\mathbf{X}) = & \beta_0 + \beta_1 \text{ASPIRIN} + \beta_2 \text{AGE} + \beta_3 \text{GENDER} \\ & + \beta_4 \text{WEIGHT} + \beta_5 \text{HEIGHT} \\ & + \beta_6 \text{ASPIRIN*AGE} + \beta_7 \text{ASPIRIN*GENDER} \\ & + \beta_8 \text{ASPIRIN*WEIGHT} \\ & + \beta_9 \text{ASPIRIN*HEIGHT} \end{aligned}$$

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.

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

Variable	Coefficient	Empirical Std Err	Wald p-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*AGE	0.0069	0.0185	0.7087
ASPIRIN*GENDER	0.9836	0.5848	0.0926
ASPIRIN*WEIGHT	-0.0147	0.0137	0.2848
ASPIRIN*HEIGHT	-0.0107	0.0218	0.6225

Odds ratio (ASPIRIN=1 vs. ASPIRIN=0)

$$\text{odds} = \exp(\beta_0 + \beta_1 \text{ASPIRIN} + \beta_2 \text{AGE} + \beta_3 \text{GENDER} + \beta_4 \text{WEIGHT} + \beta_5 \text{HEIGHT} + \beta_6 \text{ASPIRIN*AGE} + \beta_7 \text{ASPIRIN*GENDER} + \beta_8 \text{ASPIRIN*WEIGHT} + \beta_9 \text{ASPIRIN*HEIGHT})$$

Separate OR for each pattern of covariates:

$$\text{OR} = \exp(\beta_1 + \beta_6 \text{AGE} + \beta_7 \text{GENDER} + \beta_8 \text{WEIGHT} + \beta_9 \text{HEIGHT})$$

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 versus 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.

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.

This yields the expression for the *odds ratio*, also shown on the left.

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 $\exp[0.3934 + (0.0069)(60) + (0.9836)(1) + (-0.0147)(75) + (-0.0107)(170)] = 0.32$.

Chunk test

$$H_0: \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$$

Likelihood ratio test ~~X~~

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.

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 $df = \#$ 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 Chapter 11). 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 are 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)

logit $P(D=1|X) =$

$$\beta_0 + \beta_1 \text{ASPIRIN} + \beta_2 \text{AGE} + \beta_3 \text{GENDER} \\ + \beta_4 \text{WEIGHT} + \beta_5 \text{HEIGHT}$$

The no interaction model (**Model 2**) is presented at left. The GEE parameter estimates along with the working correlation matrix using the exchangeable correlation structure are also shown.

Exchangeable correlation structure

Variable	Coefficient	Empirical Std Err	Wald 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

$$\widehat{OR}_{ASPIRIN=1 \text{ vs. } ASPIRIN=0} = \exp(-1.3302) \\ = 0.264$$

Wald test

$$H_0: \beta_1 = 0$$

$$Z = \frac{-1.3302}{0.1444} = -9.21, P = 0.0001$$

Score test

$$H_0: \beta_1 = 0$$

$$\text{Chi-square} = 65.84, P = 0.0001$$

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.

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.

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

$$\hat{\rho} = -0.0954$$

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	.

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 .

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 to reocclusions from different patients.

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)

In this example, predictor values did not vary within a cluster.

A comparison of the odds ratio estimates with 95% confidence intervals for the no-interaction models of both the GEE model and SLR are 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, 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 h

$$D = \begin{cases} 1 & \text{if yes} \\ 0 & \text{if no} \end{cases}$$

Each subject has two observations

$$RX = 1$$

$$RX = 0$$

RX is time dependent: values change for each subject (cluster)

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.

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

$$\text{logit } P(D=1|\mathbf{X}) = \beta_0 + \beta_1 RX$$

$n_i = 2$: {AR1, exchangeable, or unstructured}

$$\Rightarrow \text{same } 2 \times 2 \mathbf{C}_i \quad \mathbf{C}_i = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

Exchangeable correlation structure

Variable	Coefficient	Empirical	Wald
		Std Err	p-value
INTERCEPT	-0.2007	0.3178	0.5278
RX	0.3008	0.3868	0.4368
Scale	1.0127	.	.

$$\widehat{OR} = \exp(0.3008) = 1.35$$

95% CI = (0.63, 2.88)

Exchangeable \mathbf{C}_i

	COL1	COL2
ROW1	1.0000	0.2634
ROW2	0.2634	1.0000

SLR (naive) model

Variable	Coefficient	Model-based	Wald
		Std Err	p-value
INTERCEPT	-0.2007	0.3178	0.5278
RX	0.3008	0.4486	0.5826
Scale	1.0000	.	.

$$\widehat{OR} = \exp(0.3008) = 1.35$$

95% CI = (0.56, 3.25)

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).

The output for a GEE model with an exchangeable correlation structure is presented on the left.

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.

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.

SUMMARY

- ✓ Chapter 12: 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.

Chapter 13: Other Approaches to Analysis of Correlated Data

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

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 380)**
II-IV. Examples (pages 380–395)

- 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.

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:

$$\text{logit } P(X) = \beta_0 + \beta_1 \text{BIRTHWGT} + \beta_2 \text{GENDER} + \beta_3 \text{DIARRHEA}$$

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.

Variable	Coefficient	Empirical Std Err	Model-based 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 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.

Variable	Coefficient	Empirical	
		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 Section 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

Variable	Coefficient	Empirical 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

Variable	Coefficient	Empirical Std Err
INTERCEPT	-0.5625	0.4058
RX	0.3104	0.3992
SEQUENCE	0.7118	0.5060

Model 3

Variable	Coefficient	Empirical 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 versus 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): $\exp[0.1636 \pm 1.96(0.8722)] = (0.21, 6.51)$; 95% CI (with model-based SE): $\exp[0.1636 \pm 1.96(0.2841)] = (0.67, 2.06)$.
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 are 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: $\exp[0.3601 \pm 1.96(0.8122)] = (0.29, 7.04)$.

9. Consistency is an asymptotic property. As the number of clusters approach 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 DIAR-RHEA 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.

13

Other Approaches for Analysis of Correlated Data



Contents	
Introduction	406
Abbreviated Outline	406
Objectives	407
Presentation	408
Detailed Outline	427
Practice Exercises	429
Test	433
Answers to Practice Exercises	435

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.

- I. Overview (page 408)
- II. Alternating logistic regressions algorithm (pages 409–413)
- III. Conditional logistic regression (pages 413–417)
- IV. The generalized linear mixed model approach (pages 418–425)

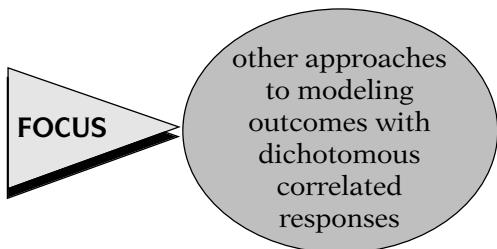
Objectives

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.

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 Chapter 8 but is presented here in a somewhat 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 (ρ 's)	odds ratios (OR's)

$$\text{OR}_{ijk} = \frac{P(Y_{ij} = 1, Y_{ik} = 1)P(Y_{ij} = 0, Y_{ik} = 0)}{P(Y_{ij} = 1, Y_{ik} = 0)P(Y_{ij} = 0, Y_{ik} = 1)}$$

GEE: α 's and β 's estimated by alternately updating estimates until convergence

ALR: α 's and β 's estimated similarly

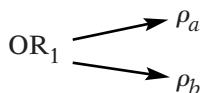
BUT

ALR: α are log OR's
(GEE: α are ρ 's)

$$\text{ALR } \text{OR}_{jk} > 1 \Leftrightarrow \text{GEE } \rho_{jk} > 1$$

$$\text{ALR } \text{OR}_{jk} < 1 \Leftrightarrow \text{GEE } \rho_{jk} < 1$$

Same OR can correspond to different ρ 's



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 j th and k th responses for the i th subject can be expressed as shown on the left.

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.

The ALR approach works in a similar manner, except that the alpha parameters are odds ratio (or log odds ratio) parameters rather than correlation parameters. Moreover, for the same data, an odds ratio between the j th and k th responses that is greater than 1 using an ALR model corresponds to a positive correlation between the j th and k th responses using a GEE model. Similarly, an odds ratio less than 1 using an ALR model corresponds to a negative correlation between responses. 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 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 Versus ALR

Aspirin–Heart Bypass Study
(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)

Model

$$\text{logit } P(\mathbf{X}) = \beta_0 + \beta_1 \text{ASPIRIN} + \beta_2 \text{AGE} + \beta_3 \text{GENDER} + \beta_4 \text{WEIGHT} + \beta_5 \text{HEIGHT}$$

The ALR model is illustrated by returning to the Aspirin–Heart Bypass Study example, which was first presented in Chapter 12. 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).

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).

Consider the model presented at left, with ASPIRIN, AGE, GENDER, WEIGHT, and HEIGHT as covariates.

EXAMPLE (continued)

GEE Approach (Exchangeable ρ)

Variable	Coefficient	Empirical Std Err	z	Wald 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			

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

ALR approach (Exchangeable OR)

Variable	Coefficient	Empirical Std Err	z	Wald 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	
<u>ALPHA1</u>	-0.4716	0.1217	0.0001	

$$\exp(\text{ALPHA1}) = \widehat{\text{OR}}_{jk}(\text{exchangeable})$$

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 Chapter 12.)

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.

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 i th subject's j th and k th 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{OR} ASPIRIN=1 vs. ASPIRIN=0 :

$$\text{GEE} \rightarrow \exp(-1.3302) = 0.264$$

$$\text{ALR} \rightarrow \exp(-1.3253) = 0.266$$

S.E. (Aspirin) = 0.1444 (GEE and ALR)

Measure of association (\widehat{OR}_{jk})

$$\widehat{OR}_{jk} = \exp(\text{ALPHA1})$$

$$= \exp(-0.4716) = 0.62$$

(Negative association:
similar to $\hat{\rho} = -0.0954$

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.

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).

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.

95% CI for ALPHA1

$$= \exp[(-0.4716 \pm 1.96(0.1217))]$$

$$= (0.49, 0.79)$$

P-value = 0.0001 \Rightarrow ALPHA1 significant

	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

ALR: allows inferences about both $\hat{\alpha}$ and $\hat{\beta}$'s

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).

III. Conditional Logistic Regression

EXAMPLE

Heartburn Relief Study
("subject" as matching factor)

40 subjects received:

- active treatment ("exposed")
- standard treatment ("unexposed")

CLR model

$$\text{logit } P(\mathbf{X}) = \beta_0 + \beta_1 RX + \sum_{i=1}^{39} \gamma_i V_i$$

where

$$V_i = \begin{cases} 1 & \text{for subject } i \\ 0 & \text{otherwise} \end{cases}$$

GEE model

$$\text{logit } P(\mathbf{X}) = \beta_0 + \beta_1 RX$$

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 Chapter 12. 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 Chapter 8, 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 fortieth subject acts as the reference group since all of the dummy variables have a value of zero for the fortieth subject (see Chapter 8).

CLR approach \Rightarrow
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 $P(\mathbf{X})$, 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

No β_0 or γ_i estimates in CLR model
(cancel out in conditional likelihood)

Returning to the Heartburn Relief Study data, the output obtained from running the conditional logistic regression is presented on the left.

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).

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).

Odds ratio and 95% CI

$$\widehat{\text{OR}} = \exp(0.4055) = 1.50$$

$$95\% \text{ CI} = (0.534, 4.214)$$

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 discussed in Chapter 12. 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 \Rightarrow parameter will not be estimated in 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)

$$\text{logit } P(\mathbf{X}) = \beta_0 + \beta_1 \text{ASPIRIN} + \beta_2 \text{AGE} \\ + \beta_3 \text{GENDER} + \beta_4 \text{WEIGHT} + \beta_5 \text{HEIGHT}$$

CLR model

Variable	Coefficient	Standard Error	Wald p-value
AGE	0	.	.
GENDER	0	.	.
WEIGHT	0	.	.
HEIGHT	0	.	.
ASPIRIN	0	.	.

All strata concordant \Rightarrow model will not run

Within-subject variability for one or more independent variable \Rightarrow

- model will run
- parameters estimated for only those variables

Matched analysis:

- Advantage: control of confounding factors
- Disadvantage: cannot separate effects of time-independent factors

Besides ASPIRIN, additional covariates include AGE, GENDER, WEIGHT, and HEIGHT. We restate the model from the previous section at left.

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 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.

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.

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**)

$$\text{logit } P(\mathbf{X}) = \beta_0 + \beta_1 RX + \sum_{i=1}^{39} \gamma_i V_i$$

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 \Rightarrow 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

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:

- special case of GLMM
- combines GEE and CLR features

GEE

User specifies
 $g(\mu)$ and \mathbf{C}_i

CLR

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 i th subject's mean response (μ_i).

EXAMPLE

Heartburn Relief Study

$$\text{logit } \mu_i = P(D=1|RX) = \beta_0 + \beta_1 RX_i + b_{0i}$$

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 i th subject's mean response, which in this case is the i th 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:

$$\text{logit of baseline risk} = (\beta_0 + b_{0i})$$

b_{0i} = subject-specific intercept

No random effect for RX \Rightarrow (assumption)
effect is same for each subject = $\exp(\beta_1)$

With this model, each subject has his/her own baseline risk, the logit of which is the intercept plus the random effect ($\beta_0 + b_{0i}$). This random effect, 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} .

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)

Variable	Coefficient	Standard Error	Wald p-value
INTERCEPT	-0.2285	0.3583	0.5274
RX	0.3445	0.4425	0.4410

Odds ratio and 95% CI:

$$\widehat{\text{OR}} = \exp(0.3445) = 1.41$$

$$95\% \text{ CI} = (0.593, 3.360)$$

The output obtained from running the MLM on the heartburn data is presented on the left. This model was run using a macro called GLIMMIX on the SAS system.

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	$\widehat{\text{OR}}$	$s_{\widehat{\beta}}$
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 at 1.35 using the GEE approach, but somewhat smaller than the estimate obtained at 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)

Random effects model:

- fixed component (fixed effects)
- random components
 - random effects
 - residual variation (error)

1. Random effects: \mathbf{b}

$$\text{Var}(\mathbf{b}) = \mathbf{G}$$

2. Residual variation: ε

$$\text{Var}(\varepsilon) = \mathbf{R}$$

Random components layered:

$$Y_{ij} = \frac{1}{1 + \exp\left[-\beta_0 + \sum_{h=1}^p \beta_h X_{hij} + \varepsilon_{1i}\right]} + \varepsilon_{2ij}$$

random
effects

residual
variation

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.

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 (\mathbf{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.

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 (ε_{1i}) and the residual variation (ε_{2ij}) for the j th response of the i th subject (Y_{ij}).

$$\varepsilon_{2ij} = Y_{ij} - P(Y_{ij}=1|\mathbf{X})$$

where

$$P(Y_{ij}=1|\mathbf{X}) = \mu$$

$$= \frac{1}{1 + \exp\left[-(\beta_0 + \sum_{h=1}^p \beta_h X_{hij} + \varepsilon_{1i})\right]}$$

The residual variation (ε_{2ij}) accounts for the difference between the observed value of Y and the $P(Y=1|\mathbf{X})$ for a given subject. The random effect (ε_{1i}), on the other hand, allows for randomness in the modeling of the mean response [i.e., $P(Y=1|\mathbf{X})$], which in contrast is modeled in a standard logistic regression solely in terms of fixed predictors (X 's) and coefficients (β 's).

	GLM	GEE
Model	$Y_i = \mu_i + \varepsilon_{2ij}$	$Y_{ij} = \mu_{ij} + \varepsilon_{2ij}$
R	Independent	Correlated
G	—	—

$$\text{GLMM: } Y_{ij} = \underbrace{g^{-1}(\mathbf{X}, \beta, \varepsilon_{1i})}_{\mu_{ij}} + \varepsilon_{2ij}$$

User specifies covariance structures for **R**, **G**, or both

For GLM and GEE models, the outcome Y is modeled as the sum of the mean and the residual variation [$Y = \mu + \varepsilon_{2ij}$, where the mean (μ) 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**).

In contrast, for GLMMs, the mean also contains a random component (ε_{1i}). 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 ~~\Rightarrow~~ unique covariance

For i th 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 = 2×2 matrix

G = 1×1 matrix (only one random effect)

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 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 versus MLM: subject-specific effects

CLR: logit $\mu_i = \beta_0 + \beta_1 RX + \gamma_i$
where γ_i is a **fixed** effect

MLM: logit $\mu_i = \beta_0 + \beta_1 RX + b_{0i}$
where b_{0i} is a **random** effect

Fixed effect γ_i : impacts modeling of μ

Random effect b_{0i} : used to characterize the variance

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 i th subject's response. Using the heartburn data, these models can be expressed as shown on the left.

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 versus MLM

GEE model: $\text{logit } \mu = \beta_0 + \beta_1 RX$

No subject-specific random effects (b_{0i})

Within-subject correlation specified in **R** matrix

MLM model: $\text{logit } \mu_i = \beta_0 + \beta_1 RX + b_{0i}$

Subject-specific random effects

A GEE model can also be expressed in terms of the i th 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 $\Rightarrow E(Y|X)$ not conditioned on cluster specific information

(e.g., not allowed as X

- earlier values of Y
- subject-specific effects)

Marginal models (examples):

GEE
ALR
SLR

A GEE model represents a type of model called a marginal model. With a marginal model, the mean response $E(Y|X)$ is not directly conditioned on any variables containing information on the within-cluster correlation. For example, the predictors (**X**) in a marginal model cannot be earlier values of the response from the same subject or subject-specific effects.

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 $\exp(\beta_1)$ 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 $\exp(\beta_1)$ depends, in part, on the type of model that is run.

Heartburn Relief Study:

GEE: marginal model

$\exp(\hat{\beta}_1)$ is population \widehat{OR}

MLM:

$\exp(\hat{\beta}_1)$ is \widehat{OR} for an *individual*

What is an individual OR?

Each subject has separate probabilities

$$P(X=1|RX=1)$$

$$P(X=1|RX=0)$$

↓

OR compares RX=1 vs. RX=0 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 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 versus 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.

<u>Goal</u>	<u>OR</u>
population inferences	\Rightarrow marginal
individual inferences	\Rightarrow 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**

Mixed models are flexible:

- layer random components
- handle nested clusters
- control for subject effects

Performance of mixed logistic models not fully evaluated

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 macro GLIMMIX, use an approach termed penalized quasi-likelihood equations (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993). Alternatively, the SAS procedure NLMIXED (nonlinear mixed) 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 the GLIMMIX macro, NLMIXED does not allow the user to specify a correlation structure for the **G** and **R** matrices (SAS Institute, 2000).

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.

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 (Chapter 11 and Chapter 12).

SUMMARY

- ✓ Chapter 13. 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 408)

- A. Other approaches to 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 409–413)

- A. Similar to GEE except that
 - i. associations between pairs of responses are modeled with odds ratios instead of correlations:

$$\text{OR}_{ijk} = \frac{P(Y_{ij} = 1, Y_{ik} = 1)P(Y_{ij} = 0, Y_{ik} = 0)}{P(Y_{ij} = 1, Y_{ik} = 0)P(Y_{ij} = 0, Y_{ik} = 1)};$$

- ii. associations between responses may also be of interest, and not considered nuisance parameters.

III. Conditional logistic regression (pages 413–417)

- 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 ($n=40$)]:

$$\text{logit } P(\mathbf{X}) = \beta_0 + \beta_1 RX + \sum_{i=1}^{39} \gamma_i V_i$$

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 418–425)

- 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:

$$\text{logit } \mu_i = P(D=1|RX) = \beta_0 + \beta_1 RX_i + b_{0i}$$

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:

- the conditional logistic model:

$$\text{logit } \mu_i = \beta_0 + \beta_1 RX + \gamma_i \quad \text{where } \gamma_i \text{ is a fixed effect}$$

- the mixed logistic model:

$$\text{logit } \mu_i = \beta_0 + \beta_1 RX + b_{0i} \quad \text{where } b_{0i} \text{ is a random effect.}$$

- E. Interpretation of the odds ratio:

- marginal model: population average OR;

- subject-specific effects model: individual OR.

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:

$$\text{Corr}(Y_j, Y_k) = \frac{AT - M_1 N_1}{\sqrt{M_1 M_0 N_1 N_0}}, \quad \text{OR} = \frac{AD}{BC}.$$

- 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

- 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?
- 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 i th subject's mean response:

$$\text{logit } \mu_i = P(D=1|RX) = \beta_0 + \beta_1 RX_i + b_{0i} + b_{1i} RX_i$$

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 of b_{0i} and b_{1i} is a 2×2 matrix, \mathbf{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:

$$\text{logit } \mu_i = P(D=1|RX) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) RX_i$$

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 Chapter 11. A covariance structure not only contains correlation parameters but variance parameters as well.

Variance components	Compound symmetric	Unstructured
$\begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$	$\begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2\rho & \sigma^2 \end{bmatrix}$	$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \sigma_1\sigma_3\rho_{13} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 & \sigma_2\sigma_3\rho_{23} \\ \sigma_1\sigma_3\rho_{13} & \sigma_2\sigma_3\rho_{23} & \sigma_3^2 \end{bmatrix}$

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		Subject 2		Pooled subjects	
	$E=1$		$E=0$		$E=1$
$D=1$	50	25	$D=1$	25	10
$D=0$	50	75	$D=0$	75	90
Total	100	100	Total	100	100
					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		Subject 2		Pooled subjects	
	E=1		E=1		E=0
D=1	50	25	D=1	20	10
D=0	50	75	D=0	80	90
Total	100	100	Total	100	100

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 i th subject.
7. Use Model 1 to estimate the odds ratios for RX (active versus standard treatment).
8. Use Model 2 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 Chapter 12)?
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 $T = A + D$.

$$\text{corr} = \frac{AT - M_1 N_1}{\sqrt{M_1 M_0 N_1 N_0}} = \frac{A(A + D) - AD}{\sqrt{(AD)(AD)}} = 1.$$

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 $T = B + C$.

$$\text{corr} = \frac{AT - M_1 N_1}{\sqrt{M_1 M_0 N_1 N_0}} = \frac{0(B + C) - BC}{\sqrt{(BC)(BC)}} = -1.$$

The corresponding odds ratio is zero.

3. If $AD = BC$, then $D = (BC)/A$ and $T = A + B + C + (BC)/A$.

$$\text{corr} = \frac{AT - M_1 N_1}{\sqrt{M_1 M_0 N_1 N_0}} = \frac{A[A + B + C + (BC/A)] - [(A+B)(A+C)]}{\sqrt{M_1 M_0 N_1 N_0}} = 0$$

$OR = \frac{AD}{BC} = \frac{AD}{AD} = 1$ (indicating no association between Y_j and Y_k).

$$\frac{\exp(\beta_1 + b_{1i})}{1 + \exp[-(\beta_0 + b_{0i})]}$$

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.

A

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 8.2, SPSS version 10.0, and Stata version 7.0. 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 four 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 8.0 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

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 Chapter 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 diastolic blood pressure is greater than or equal to 160 or the systolic blood pressure is greater than or equal to 95.
CH, CC	Product terms of CAT \times HPT and CAT \times CHL, respectively.

MI Dataset (mi.dat)

This dataset is used to demonstrate conditional logistic regression. The MI dataset is discussed in Chapter 8. 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 and 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.

Cancer Dataset (cancer.dat)

This dataset is used to demonstrate polytomous and ordinal logistic regression. The cancer dataset, discussed in Chapters 9 and 10, 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 in one of these age groups.
SMK	A dichotomous variable indicating whether the subject is (coded 1) or is not (coded 0) a current smoker.

Infant Dataset (infant.dat)

This is the dataset used to demonstrate GEE modeling. The infant dataset, discussed in Chapters 11 and 12, is part of a health intervention study in Brazil (Cannon et al., 2001). The study involves 168 infants, each of whom has at least 5 and up to 9 monthly measurements, yielding 1458 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 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 -1 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.

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 four SAS procedures are used to perform the analyses in this appendix.

1. **PROC LOGISTIC:** This procedure can be used to run logistic regression and ordinal logistic regression using the proportional odds model.
2. **PROC GENMOD:** This procedure can be used to run GLM (including logistic regression and ordinal logistic regression) and GEE models.
3. **PROC CATMOD:** This procedure can be used to run polytomous logistic regression.
4. **PROC PHREG:** This procedure can be used to run conditional logistic regression.

The capabilities of these procedures are not limited to performing the above-listed analyses. For example, PROC PHREG also may be used to run Cox proportional hazard models for survival analyses. However, our goal is to demonstrate only the types of modeling presented in this text.

Unconditional Logistic Regression

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 \times HPT; CC is the product CAT \times CHL. The variables representing the interaction terms have already been included in the datasets provided on the accompanying disk.

The model is stated as follows:

$$\text{logit } P(\text{CHD}=1 | \mathbf{X}) = \beta_0 + \beta_1 \text{CAT} + \beta_2 \text{AGE} + \beta_3 \text{CHL} + \beta_4 \text{ECG} + \beta_5 \text{SMK} + \beta_6 \text{HPT} + \beta_7 \text{CH} + \beta_8 \text{CC}$$

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 that the file is located on the A drive (i.e., on a disk). 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 'A:\';
```

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 lowercase letters, SAS reads them as uppercase. 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 for the PROC LOGISTIC. However, a temporary SAS dataset has to be re-created 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 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

Criterion	Intercept Only	Intercept and Covariates
AIC	440.558	365.230
SC	444.970	404.936
-2 Log L	438.558	347.230

Analysis of Maximum Likelihood Estimates

Parameter	DF	Standard 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

Effect	Odds Ratio Estimates		
	Point Estimate	95% Wald 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.

PROC GENMOD

Next, we illustrate the use of PROC GENMOD with the Evans County data. PROC GENMOD can be used to run GLM and 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 'LOG OR (CHL=220, HPT=1)' CAT 1 CC 220 CH 1/EXP;
ESTIMATE 'LOG 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. However, this may not be consistent with other SAS versions (e.g., CHD=1 is the default event for PROC GENMOD in version 8.0 but not in version 8.2). 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 Value	chd	Total 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

Parameter	Estimate	Standard Error	Wald 95% Confidence Limits	Chi-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

Label	Estimate	Standard Error	Confidence Limits	Chi-Square	Pr > ChiSq
Log OR (chl=220, hpt=1)	0.1960	0.4774	-0.7397 1.1318	0.17	0.6814
Exp(Log OR (chl=220, hpt=1))	1.2166	0.5808	0.4772 3.1012		
Log OR (chl=220, hpt=0)	2.5278	0.6286	1.2957 3.7599	16.17	<.0001
Exp(Log OR (chl=220, hpt=0))	12.5262	7.8743	3.6537 42.9445		

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.

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 8 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:

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 ≥ 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
;
RUN;
```

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 DECSENDING option is not necessary if the response is in the EVENTS / TRIALS form. The output is omitted.

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 \times CAT \times AGEGRP \times 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
;
RUN;
```

Whereas the dataset EVANS2 contains 8 observations, the dataset EVANS3 contains 16 observations. 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, whereas 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;
```

THE ANALYST APPLICATION

The procedures described above are run by entering the appropriate code in the Program (or Enhanced) Editor window and then submitting 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

PROC PHREG

Next, a conditional logistic regression is demonstrated with the MI dataset using PROC PHREG. 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:

$$\text{logit } P(\text{CHD} = 1 | \mathbf{X}) = \beta_0 + \beta_1 \text{SMK} + \beta_2 \text{SPB} + \beta_3 \text{ECG} + \sum_{i=1}^{38} \gamma_i V_i$$

$$V_i = \begin{cases} 1 & \text{if } i\text{th matched triplet} \\ 0 & \text{otherwise} \end{cases} \quad i = 1, 2, \dots, 38$$

PROC PHREG is a procedure developed for survival analysis, but it can also be used to run a conditional logistic regression. In order to apply PROC PHREG for a conditional logistic regression, a *time* variable must be created in the data even though a time variable is not defined for use in the study. The *time* variable should be coded to indicate that all cases had the event at the same time and all controls were censored at a later time. This can easily be accomplished by defining a variable that has the value 1 for all the cases and the value 2 for all the controls. Do not look for meaning in this variable, as its purpose is just a tool to get the computer to perform a conditional logistic regression.

With the MI data, we shall use the permanent SAS dataset (**mi.sas7bdat**) to create a new SAS dataset that creates the time variable (called SURVTIME) needed to run the conditional logistic regression. The code follows:

```
LIBNAME REF 'A:\';

DATA MI;
SET REF.MI;
IF MI = 1 THEN SURVTIME = 1;
IF MI = 0 THEN SURVTIME = 2;
RUN;
```

A temporary SAS dataset named MI has been created that contains a variable SURVTIME along with the other variables that were contained in the permanent SAS dataset stored on the A drive. If the permanent SAS dataset was stored in a different location, the LIBNAME statement would have to be adjusted to point to that location.

The SAS procedure, PROC PRINT, can be used to view the MI dataset in the output window:

```
PROC PRINT DATA = MI; RUN;
```

The output for the first nine observations from running the PROC PRINT follows:

Obs	MATCH	PERSON	MI	SMK	SBP	ECG	SURVTIME
1	1	1	1	0	160	1	1
2	1	2	0	0	140	0	2
3	1	3	0	0	120	0	2
4	2	4	1	0	160	1	1
5	2	5	0	0	140	0	2
6	2	6	0	0	120	0	2
7	3	7	1	0	160	0	1
8	3	8	0	0	140	0	2
9	3	9	0	0	120	0	2

The code to run the conditional logistic regression follows:

```
PROC PHREG DATA=MI;
MODEL SURVTIME*MI(0) = SMK SBP ECG / TIES=DISCRETE;
STRATA MATCH;
RUN;
```

The model statement contains the time variable (SURVTIME) followed by an asterisk and the case status variable (MI) with the value of the noncases (0) in parentheses. The TIES = DISCRETE option in the model statement requests that the conditional logistic likelihood be applied, assuming that all the cases have the same value of SURVTIME. The PROC PHREG output follows:

The PHREG Procedure

Model Information

Data Set	WORK.MI
Dependent Variable	survtime
Censoring Variable	mi
Censoring Value(s)	0
Ties Handling	DISCRETE

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	85.692	63.491
AIC	85.692	69.491
SBC	85.692	74.482

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	22.2008	3	<.0001
Score	19.6785	3	0.0002
Wald	13.6751	3	0.0034

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
SMK	1	0.72906	0.56126	1.6873	0.1940	2.073
SBP	1	0.04564	0.01525	8.9612	0.0028	1.047
ECG	1	1.59926	0.85341	3.5117	0.0609	4.949

Although the output calls it a hazard ratio, the *odds ratio* estimate for SMK=1 vs. SMK=0 is $\exp(0.72906) = 2.073$.

ANALYST APPLICATION

As with PROC LOGISTIC, PROC PHREG can also be invoked in the Analyst Application. The first step is to open the SAS dataset **mi.sas7bdat** in the Analyst spreadsheet. To perform a conditional logistic regression, select Statistics → Survival → Proportional Hazards. In the dialog box, select SURVTIME as the Time variable. Select MI as the Censoring variable and indicate that the Censoring Value for the data is 0 (i.e., the value for the controls is 0). Next, input SMK, SBP, and ECG into the Explanatory box. Under the Methods button, select “Discrete logistic model” as the “Method to handle failure time ties.” (Breslow is the default setting.) Under the Variables button, select MATCH as the Strata variable. Click on OK in the main dialog box to run the model. The output generated is from PROC PHREG. It is similar to the output presented previously and is thus omitted here. Again, the Log window in SAS can be viewed to see the code for the procedure.

Polytomous Logistic Regression

Next, a polytomous logistic regression is demonstrated with the cancer dataset using PROC CATMOD. If the permanent SAS dataset **cancer.sas7bdat** is in the A 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 'A:\';
```

First, a PROC PRINT will be 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 CATMOD is the SAS procedure that is used to run a polytomous logistic regression. (This procedure is not available in the Analyst Application.)

The three-category outcome variable is SUBTYPE, coded as 0 for Adeno-squamous, 1 for Adenocarcinoma, and 2 for Other. The model is stated as follows:

$$\ln \left[\frac{P(\text{SUBTYPE} = g | \mathbf{X})}{P(\text{SUBTYPE} = 0 | \mathbf{X})} \right] = \alpha_g + \beta_{g1}\text{AGE} + \beta_{g2}\text{ESTROGEN} + \beta_{g3}\text{SMOKING}$$

where $g = 1, 2$.

By default, PROC CATMOD assumes the highest level of the outcome variable is the reference group, as does PROC LOGISTIC. Unfortunately, PROC CATMOD does not have a DESCENDING option, as does PROC LOGISTIC. If we wish to make $\text{SUBTYPE}=0$ (i.e., Adenosquamous), the reference group using PROC CATMOD, the variable SUBTYPE must first be sorted in descending order using PROC SORT and then the ORDER = DATA option must be used in PROC CATMOD. The code follows:

```
PROC SORT DATA = REF.CANCER OUT = CANCER;
BY DESCENDING SUBTYPE;
RUN;

PROC CATMOD ORDER = DATA DATA = CANCER;
DIRECT AGE ESTROGEN SMOKING;
MODEL SUBTYPE = AGE ESTROGEN SMOKING;
RUN;
```

PROC CATMOD treats all independent variables as nominal variables and recodes them as dummy variables by default. The DIRECT statement in PROC CATMOD, followed by a list of variables, is used if recoding of those variables is not desired.

The PROC CATMOD output follows:

The CATMOD Procedure

Response	subtype	Response Levels	3
Weight Variable	None	Populations	8
Data Set	CANCER	Total Frequency	286
Frequency Missing	2	Observations	286

Response Profiles					
	Response	subtype			
	1	2			
	2	1			
	3	0			

Analysis of Maximum Likelihood Estimates					
Effect	Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-1.2032	0.3190	14.23	0.0002
	2	-1.8822	0.4025	21.87	<.0001
AGE	3	0.2823	0.3280	0.74	0.3894
	4	0.9871	0.4118	5.75	0.0165
ESTROGEN	5	-0.1071	0.3067	0.12	0.7270
	6	-0.6439	0.3436	3.51	0.0609
SMOKING	7	-1.7913	1.0460	2.93	0.0868
	8	0.8895	0.5254	2.87	0.0904

Notice that 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 $\exp(0.2823) = 1.33$.

Ordinal Logistic Regression

PROC LOGISTIC

Next, an 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:

$$\ln \left[\frac{P(\text{GRADE} \geq g | \mathbf{X})}{P(\text{GRADE} < g | \mathbf{X})} \right] = \alpha_g + \beta_1 \text{AGE} + \beta_2 \text{ESTROGEN}$$

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 Value	grade	Total 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
RACE	1	0.4270	0.2720	2.4637	0.1165
ESTROGEN	1	-0.7763	0.2493	9.6954	0.0018

Effect	Odds Ratio Estimates		
	Point Estimate	95% Confidence Limits	Wald
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.

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, whereas the response variable follows the multinomial distribution.

The output is omitted.

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 Data with GEE

Next, the programming of a GEE model with the infant care dataset is demonstrated using PROC GENMOD. The model is stated as follows:

$$\text{logit } P(\text{OUTCOME} = 1 | \mathbf{X}) = \beta_0 + \beta_1 \text{BIRTHWGT} + \beta_2 \text{GENDER} + \beta_3 \text{DIARRHEA}.$$

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 A drive.

```
LIBNAME REF 'A:\';
PROC PRINT DATA = REF.INFANT; RUN;
```

The output for the first nine observations obtained from running the PROC PRINT is presented. These are data on just one infant. Each observation represents one of the nine monthly measurements.

Obs	IDNO	MONTH	OUTCOME	BIRTHWGT	GENDER	DIARRHEA
1	00001	1	0	3000	1	0
2	00001	2	0	3000	1	0
3	00001	3	0	3000	1	0
4	00001	4	0	3000	1	1
5	00001	5	0	3000	1	0
6	00001	6	0	3000	1	0
7	00001	7	0	3000	1	0
8	00001	8	0	3000	1	0
9	00001	9	0	3000	1	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 of 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 option 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., $\text{var}(Y) = \phi\mu(1-\mu)$].

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 Value	outcome	Total 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

Parameter	DF	Estimate	Standard Error	Wald Limits	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-1.4362	0.6022	-2.6165	-0.2559	5.69	0.0171
BIRTHWT	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
BIRTHWT	-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

Label	Estimate	Standard Error	95% Confidence Limits	Chi-Square	Pr > ChiSq
log odds ratio (DIARRHEA 1 vs 0)	0.2214	0.8558	-1.4559 1.8988	0.07	0.7958
Exp(log odds ratio (DIARRHEA 1 vs 0))	1.2479	1.0679	0.2332 6.6779		

Contrast Results for GEE Analysis

Contrast	DF	Chi-Square	Pr > ChiSq	Type
Score Test BIRTHWT 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. 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 Chapter 13, 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 1000 iterations.

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 three 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 <i>one case</i> per stratum, with one (or more) controls.

SPSS does not perform GEE logistic regression for correlated data in version 10.0.

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 \times HPT, whereas CC is the product: CAT \times CHL. The variables representing the interaction terms have already been included in the SPSS dataset **evans.sav**.

The model is restated as follows:

$$\text{logit } P(\text{CHD} = 1 | \mathbf{X}) = \beta_0 + \beta_1 \text{CAT} + \beta_2 \text{AGE} + \beta_3 \text{CHL} + \beta_4 \text{ECG} + \beta_5 \text{SMK} \\ + \beta_6 \text{HPT} + \beta_7 \text{CH} + \beta_8 \text{CC}$$

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 a disk is

```
GET
FILE='A:\evans.sav'.
```

There are two procedures which can be used to fit a standard (binary) logistic regression model: LOGISTIC REGRESSION and NOMREG (Multinomial Logistic Regression). The LOGISTIC REGRESSION procedure performs a standard logistic regression for a dichotomous outcome, whereas the NOMREG procedure can be used for dichotomous or polytomous outcomes.

To run the LOGISTIC REGRESSION procedure, select Analyze \rightarrow Regression \rightarrow 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 chl 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 follows:

```
LOGISTIC REGRESSION VAR=chd
/METHOD=ENTER cat age chl 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

<u>Unweighted cases^a</u>		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

<u>Original Value</u>	<u>Internal Value</u>
.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

Step	CAT	B	S.E.	Wald	df	Sig.	Exp(B)	95.0%	C.I. for	Lower	Upper
								EXP(B)			
1 ^a	CAT	-12.688	3.104	16.705	1	.000	.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 assess 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.

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 0 corresponds to the CHD value of 1.

The output is omitted.

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. The process is similar to that demonstrated in the SAS section with PROC PHREG and the MI dataset, although SAS is not limited to the special case.

Recall that the MI dataset contains information on 39 cases diagnosed with myocardial infarction, each of which is matched with 2 controls. Thus, it meets the criterion of one case per stratum. As with SAS, 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. In the SAS example, this variable was named SURVTIME. 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='A:\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, is

```
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

Next, a 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:

$$\ln \left[\frac{P(\text{SUBTYPE} = g | \mathbf{X})}{P(\text{SUBTYPE} = 0 | \mathbf{X})} \right] = \alpha_g + \beta_{g1}\text{AGE} + \beta_{g2}\text{ESTROGEN} + \beta_{g3}\text{SMOKING}$$

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 Chapter 9, 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)	
			Lower Bound	Upper Bound
NEWTYPE				
.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

The final analysis shown is ordinal logistic regression 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:

$$\ln \left[\frac{P(\text{GRADE} \leq g^* | \mathbf{X})}{P(\text{GRADE} > g^* | \mathbf{X})} \right] = \alpha_{g^*}^* - \beta_1^* \text{AGE} - \beta_2^* \text{ESTROGEN} \quad \text{for } g^* = 0, 1.$$

Note that this is the alternative formulation of the ordinal model discussed in Chapter 9. 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 i th 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 function similar to 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 .
```

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.
a Link function: Logit.

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 ($P = 0.638$), indicating 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.

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.

logit This command is used to run logistic regression

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.

clogit This command is used to run conditional logistic regression.

mlogit This command is used to run polytomous logistic regression.

ologit This command is used to run ordinal logistic regression.

xtgee This command is used to run GEE models.

lrtest This command is used to perform likelihood ratio tests.

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. Command 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 \times HPT; CC is the product CAT \times CHL. The variables representing the interaction terms have already been included in the Stata dataset **evans.dta**.

The model is restated as follows:

$$\text{logit } P(\text{CHD} = 1|\mathbf{X}) = \beta_0 + \beta_1 \text{CAT} + \beta_2 \text{AGE} + \beta_3 \text{CHL} + \beta_4 \text{ECG} + \beta_5 \text{SMK} + \beta_6 \text{HPT} + \beta_7 \text{CH} + \beta_8 \text{CC}.$$

The first step is to activate the Evans dataset by clicking on File \rightarrow 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		
chl	-.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 eight regression parameters to a reduced model containing only the intercept. The test statistic follows a chi-square distribution with eight 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 typing the following command:

```
lrtest, saving(0)
```

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
```

The resulting output follows:

Logit: likelihood-ratio test	chi2(2)	=	53.16
	Prob > chi2	=	0.0000

The chi-square statistic with two degrees of freedom is 53.16, which is statistically significant as the *P*-value is close to 0.

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 are 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:

$$\text{logit } P(\text{CHD} = 1) = \beta_0 + \beta_1 \text{CAT} + \beta_2 \text{AGEGRP} + \beta_3 \text{ECG}$$

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 16 observations rather than 8. 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 as follows:

```
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.

Conditional Logistic Regression

Next, a 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:

$$\text{logit } P(\text{CHD} = 1 | \mathbf{X}) = \beta_0 + \beta_1 \text{SMK} + \beta_2 \text{SPB} + \beta_3 \text{ECG} + \sum_{i=1}^{38} \gamma_i V_i$$

$$V_i = \begin{cases} 1 & \text{if } i\text{th matched triplet} \\ 0 & \text{otherwise} \end{cases} \quad i = 1, 2, \dots, 38$$

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
				Pseudo R2	=	0.2591
Log likelihood = -31.745464						
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

Next, a 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:

$$\ln \left[\frac{P(\text{SUBTYPE} = g | \mathbf{X})}{P(\text{SUBTYPE} = 0 | \mathbf{X})} \right] = \alpha_g + \beta_{g1}\text{AGE} + \beta_{g2}\text{ESTROGEN} + \beta_{g3}\text{SMOKING},$$

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:

Ordinal Logistic Regression

Next, an 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:

$$\ln \left[\frac{P(\text{GRADE} \leq g^* | \mathbf{X})}{P(\text{GRADE} > g^* | \mathbf{X})} \right] = \alpha_{g^*}^* - \beta_1^* \text{AGE} - \beta_2^* \text{ESTROGEN} \quad \text{for } g^* = 0, 1.$$

This is the alternative formulation of the proportional odds model discussed in Chapter 9. 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 i th 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						
Log likelihood = -287.60598						Number of obs = 286
						LR chi2(2) = 19.71
						Prob > chi2 = 0.0001
						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 Dichotomous Data with GEE

Finally, a GEE model is demonstrated in Stata with the infant care dataset ([infant.dta](#)). GEE models are 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) i(idno) t(month) robust
```

Following the command **xtgee** comes 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 **i()** option specifies the cluster variable and the **t()** option specifies the time the observation was made within the cluster. The robust option requests empirical-based standard errors. The options **corr(ind)**, **corr(exc)**, **corr(sta 4)**, 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)

outcome	Coef.	Std. Err.	Semi-robust			[95% Conf. Interval]
			z	P> z		
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(ar 1) i(idno) t(month) 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)

outcome	Coef.	Std. Err.	Semi-robust			[95% Conf. Interval]
			z	P> z		
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

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. $P(\mathbf{X}) = 1/(1 + \exp\{-[\alpha + \beta_1(\text{AGE}) + \beta_2(\text{SMK}) + \beta_3(\text{SEX}) + \beta_4(\text{CHOL}) + \beta_5(\text{OCC})]\})$.
23. $P(\mathbf{X}) = 1/(1 + \exp\{-[-4.32 + 0.0274(\text{AGE}) + 0.5859(\text{SMK}) + 1.1523(\text{SEX}) + 0.0087(\text{CHOL}) - 0.5309(\text{OCC})]\})$.
24. $\text{logit } P(\mathbf{X}) = -4.32 + 0.0274(\text{AGE}) + 0.5859(\text{SMK}) + 1.1523(\text{SEX}) + 0.0087(\text{CHOL}) - 0.5309(\text{OCC})$.
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

$$\begin{aligned} P(\mathbf{X}) &= 1/(1 + \exp\{-[-4.32 + 0.0274(40) + 0.5859(1) + 1.1523(1) + 0.0087(200) - 0.5309(1)]\}) \\ &= 1/(1 + \exp[-(-0.2767)]) \\ &= 1/(1 + 1.319) \\ &= 0.431. \end{aligned}$$

26. For a 40-year-old male *nonsmoker* with CHOL = 200 and OCC = 1,
 $\mathbf{X} = (\text{AGE} = 40, \text{SMK} = 0, \text{SEX} = 1, \text{CHOL} = 200, \text{OCC} = 1)$
and

$$\begin{aligned}\hat{P}(\mathbf{X}) &= 1/(1 + \exp\{-[-4.32 + 0.0274(40) + 0.5859(0) + 1.1523(1) \\&\quad + 0.0087(200) - 0.5309(1)]\}) \\&= 1/[1 + \exp[-(-0.8626)]] \\&= 1/(1 + 2.369) \\&= 0.297\end{aligned}$$

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. $\widehat{\text{OR}}(\text{SMK} \text{ controlling for AGE, SEX, CHOL, OCC})$
 $= e^{\hat{\beta}}$ 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. $\widehat{\text{OR}}(\text{OCC} \text{ controlling for AGE, SEX, SMK, CHOL})$
 $= e^{\hat{\beta}}$, where $\hat{\beta} = -0.5309$ 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. F: the number of dummy variables should be 19
10. F: $OR = \exp(\beta + \delta_1 OBS + \delta_2 PAR)$

11. The model in logit form is given as follows:

$$\text{logit } P(\mathbf{X}) = \alpha + \beta \text{CON} + \gamma_1 \text{PAR} + \gamma_2 \text{NP} + \gamma_3 \text{ASCM} + \delta_1 \text{CON} \times \text{PAR} + \delta_2 \text{CON} \times \text{NP} + \delta_3 \text{CON} \times \text{ASCM}.$$
12. The odds ratio expression is given by

$$\exp(\beta + \delta_1 \text{PAR} + \delta_2 \text{NP} + \delta_3 \text{ASCM}).$$
13. The model for the matched pairs case-control design is given by

$$\text{logit } P(\mathbf{X}) = \alpha + \beta \text{CON} + \sum_{i=1}^{199} \gamma_i V_i + \gamma_{200} \text{PAR} + \delta \text{CON} \times \text{PAR},$$
where the V_i are dummy variables which indicate the matching strata.
14. The risk for an exposed person in the first matched pair with PAR = 1 is

$$R = \frac{1}{1 + \exp[-(\alpha + \beta + \gamma_1 + \gamma_{200} + \delta \text{PAR})]}.$$
15. The odds ratio expression for the matched pairs case-control model is

$$\exp(\beta + \delta \text{PAR})$$

Chapter 3

1. a. $ROR = \exp(\beta)$
 b. $ROR = \exp(5\beta)$
 c. $ROR = \exp(2\beta)$
 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. $ROR = \exp(\beta + \delta_1AGE + \delta_2CHL)$
 b. $ROR = \exp(5\beta + 5\delta_1AGE + 5\delta_2CHL)$
 c. $ROR = \exp(2\beta + 2\delta_1AGE + 2\delta_2CHL)$
 d. For a given specification of AGE and CHL, all three estimated odds ratio 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_1 's and the three δ_2 's.
3. a. $ROR = \exp(5\beta + 5\delta_1AGE + 5\delta_2SEX)$
 b. $ROR = \exp(\beta + \delta_1AGE + \delta_2SEX)$
 c. $ROR = \exp(\beta + \delta_1AGE + \delta_2SEX)$
 d. For a given specification of AGE and SEX, the odds ratios in parts b and c should have the same value.
4. a. $\text{logit } P(\mathbf{X}) = \alpha + \beta_1S_1 + \beta_2S_2 + \gamma_1AGE + \gamma_2SEX$, 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

$$ROR = \exp(-\beta_1)$$
, where $\mathbf{X}^* = (0, 0, AGE, SEX)$
 and $\mathbf{X}^{**} = (1, 0, AGE, SEX)$.
 c. $\text{logit } P(\mathbf{X}) = \alpha + \beta_1S_1 + \beta_2S_2 + \gamma_1AGE + \gamma_2SEX + \delta_1(S_1 \times AGE) + \delta_2(S_1 \times SEX) + \delta_3(S_2 \times AGE) + \delta_4(S_2 \times SEX)$
 d. $ROR = \exp(-\beta_1 - \delta_1AGE - \delta_2SEX)$
5. a. $ROR = \exp(10\beta_3)$
 b. $ROR = \exp(195\beta_1 + 10\beta_3)$
6. a. $ROR = \exp(10\beta_3 + 10\delta_{31}AGE + 10\delta_{32}RACE)$
 b. $ROR = \exp(195\beta_1 + 10\beta_3 + 195\delta_{11}AGE + 195\delta_{12}RACE + 10\delta_{31}AGE + 10\delta_{32}RACE)$

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 somewhat 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. $H_0: \beta_{SMK} = 0$ 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 $Z = \frac{\hat{\beta}_{SMK}}{S_{\hat{\beta}_{SMK}}}$, 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 ;

$$\text{test computation: } Z = \frac{1.4361}{0.3167} = 4.53; \text{ alternatively, } Z^2 = 20.56;$$

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

$$\begin{aligned} \exp\left[\hat{\beta}_{\text{SMK}} \pm 1.96 \sqrt{\text{Var}(\hat{\beta}_{\text{SMK}})}\right] &= \exp(1.4361 \pm 1.96 \times 0.3617) \\ &= (e^{0.7272}, e^{2.1450}) = (2.07, 8.54). \end{aligned}$$

5. Null hypothesis for the likelihood ratio test for the effect of $\text{SMK} \times \text{NS}$: $H_0: \beta_{\text{SMK} \times \text{NS}} = 0$ where $\beta_{\text{SMK} \times \text{NS}}$ is the coefficient of $\text{SMK} \times \text{NS}$ in model II;

Likelihood ratio statistic: $\text{LR} = -2 \ln \hat{L}_I - (-2 \ln \hat{L}_{II})$ where \hat{L}_I and \hat{L}_{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 $\text{SMK} \times \text{NS}$ 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 $\text{SMK} \times \text{NS}$ in model II;

Wald statistic: $Z = \frac{\hat{\beta}_{\text{SMK} \times \text{NS}}}{s_{\hat{\beta}_{\text{SMK} \times \text{NS}}}}$, 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 ;

$$\text{test computation: } Z = \frac{-1.1128}{0.5997} = -1.856; \text{ alternatively, } Z^2 = 3.44;$$

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

$$\widehat{OR}_{\text{adj}} = \exp(\hat{\beta}_{\text{SMK}} + \hat{\delta}_{\text{SMK} \times \text{NS}} \text{NS}) = \exp(1.9381 - 1.1128 \text{NS})$$

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

$$\text{NS} = 1: \exp[1.9381 - 1.1128(1)] = \exp(0.8253) = 2.28;$$

$$\text{NS} = 0: \exp[1.9381 - 1.1128(0)] = \exp(1.9381) = 6.95$$

9. Formula for the 95% confidence interval for the adjusted odds ratio when NS = 1:

$$\exp\left[\hat{l} \pm 1.96\sqrt{\widehat{\text{var}}(\hat{l})}\right], \text{ where } \hat{l} = \hat{\beta}_{\text{SMK}} + \hat{\delta}_{\text{SMK} \times \text{NS}}(1) = \hat{\beta}_{\text{SMK}} + \hat{\delta}_{\text{SMK} \times \text{NS}}$$

and

$$\widehat{\text{var}}(\hat{l}) = \widehat{\text{var}}(\hat{\beta}_{\text{SMK}}) + (1)^2 \widehat{\text{var}}(\hat{\delta}_{\text{SMK} \times \text{NS}}) + 2(1) \widehat{\text{cov}}(\hat{\beta}_{\text{SMK}}, \hat{\delta}_{\text{SMK} \times \text{NS}}),$$

where $\widehat{\text{var}}(\hat{\beta}_{\text{SMK}})$, $\widehat{\text{var}}(\hat{\delta}_{\text{SMK} \times \text{NS}})$, and $\widehat{\text{cov}}(\hat{\beta}_{\text{SMK}}, \hat{\delta}_{\text{SMK} \times \text{NS}})$ are obtained from the printout of the variance–covariance matrix.

$$10. \hat{l} = \hat{\beta}_{\text{SMK}} + \hat{\delta}_{\text{SMK} \times \text{NS}} = 1.9381 + (-1.1128) = 0.8253$$

$$\widehat{\text{var}}(\hat{l}) = 0.1859 + (1)^2 (0.3596) + 2(1)(-0.1746) = 0.1859 + 0.3596 - 0.3492 = 0.1963.$$

The 95% confidence interval for the adjusted odds ratio is given by

$$\exp\left[\hat{l} \pm 1.96\sqrt{\widehat{\text{Var}}(\hat{l})}\right] = \exp(0.8253 \pm 1.96\sqrt{0.1963}) = \exp(0.8253 \pm 1.96 \times 0.4430)$$

$$= (e^{-0.0430}, e^{1.6936}) = (0.96, 5.44).$$

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 $SMK \times NS \times AS$, then the lower-order components are SMK , NS , AS , $SMK \times NS$, $SMK \times AS$, and $NS \times AS$; all these lower-order components are contained in the model.
12. A test for the term $SMK \times NS \times AS$ is not dependent on the coding of SMK because the model is hierarchically well formulated and $SMK \times NS \times AS$ is the highest-order term in the model.
13. A test for the terms $SMK \times NS$ 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 $SMK \times NS \times AS$, then test for significance of lower-order interactions $SMK \times NS$ and $SMK \times AS$, 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 $SMK \times NS \times AS$ is significant, then $SMK \times NS$ and $SMK \times AS$ are interaction terms that must remain in any further model considered. The V variables that must remain in further models are NS , AS , $NS \times AS$, 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 $NS \times AS$ are lower components of $SMK \times NS \times AS$ and because the V^* variables are matching variables.

Chapter 7

1. The interaction terms are $SMK \times NS$, $SMK \times AS$, and $SMK \times NS \times AS$. The product term $NS \times AS$ 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, $SMK \times NS \times AS$. 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 $SMK \times NS \times AS$ is significant. Finally, without doing statistical testing, the V variables need to be assessed for confounding and precision.

3. If $SMK \times NS$ is the only interaction found significant, then the model remaining after interaction assessment contains the V^* terms, SMK , NS , AS , $NS \times AS$, and $SMK \times NS$. The variable NS cannot be deleted from any further model considered because it is a lower-order component of the significant interaction term $SMK \times NS$. 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 $\exp(\beta + \delta_1 NS)$.
5. The odds ratio expression for the model that does not contain $NS \times AS$ 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 $NS \times AS$ 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 $NS \times AS$ cannot be dropped and is considered a confounder. If the odds ratio does not change, then $NS \times AS$ 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 $NS \times AS$. If the latter confidence interval is meaningfully narrower, then precision is gained by dropping $NS \times AS$, so that this variable should, therefore, be dropped. Otherwise, one should control for $NS \times AS$ 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 $NS \times AS$ 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 $NS \times AS$) 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**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. $\text{logit } P(\mathbf{X}) = \alpha + \beta E + \sum_{i=1}^{8501} \gamma_{1i} V_{1i}$,
where the V_{1i} denote 8501 dummy variables used to indicate the 8502 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

$$\exp\left[\hat{\beta} \pm 1.96 \sqrt{\widehat{\text{var}}(\hat{\beta})}\right]$$
,
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.
11. Conditional ML estimation was used to fit the model because the study design involved matching; the number of parameters in the model, including dummy variables for matching, is large relative to the number of observations in the study.
12. The variables 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 (corresponding to dummy variables) which are not estimated by the conditional ML estimation method.
13. $\text{logit } P(\mathbf{X}) = \alpha + \beta \text{SMK} + \sum_i \gamma_{1i} V_{1i} + \gamma_{21} \text{NS} + \gamma_{22} \text{AS} + \delta \text{SMK} \times \text{NS}$,
where V_{1i} denote dummy variables (the number not specified) used to indicate matching strata.

14. The squared Wald statistic (a chi-square statistic) for the product term $SMK \times NS$ is computed to be 3.44, which has a P -value of 0.0635. This is not significant at the 5% level but is significant at the 10% level. This is not significant using the traditional 5% level, but it is close enough to 5% to suggest possible interaction. Information for computing the likelihood ratio test is not provided here, although the likelihood ratio test would be more appropriate to use in this situation. (Note, however, that the LR statistic is 3.51, which is also not significant at the 0.05 level, but significant at the 0.10 level.)
15. The formula for the estimated odds ratio is given by

$$\widehat{OR}_{adj} = \exp(\hat{\beta}_{SMK} + \hat{\delta}_{SMK \times NS} NS) = \exp(1.9381 - 1.1128NS),$$
where the coefficients come from the printout and the confounding effects of NS and AS are controlled.
16. Using the adjusted odds ratio formula given in Question 8, the estimated odds ratio values for $NS = 1$ and $NS = 0$ are
 $NS = 1: \exp[1.9381 - 1.1128(1)] = \exp[0.8253] = 2.28,$
 $NS = 0: \exp[1.9381 - 1.1128(0)] = \exp[1.9381] = 6.95.$
17. When $NS = 1$, the point estimate of 2.28 shows a moderate effect of smoking, i.e., the risk for smokers is about 2.3 times the risk for non-smokers. However, the confidence interval of (0.96, 5.44) is very wide and contains the null value of 1. This indicates that the point estimate is strongly nonsignificant and is highly unreliable. Conclude that, for $NS = 1$, there is no significant evidence to indicate that smoking is related to cervical cancer.
18. Using the printout for the no interaction model, the estimated odds ratio adjusted for the other variables in model 1 is given by $e^{1.4361} = 4.20$.

The Wald statistic $Z = \frac{\hat{\beta}_{SMK}}{S_{\hat{\beta}_{SMK}}}$ 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.

The 95% interval estimate for the above odds ratio is given by

$$\begin{aligned} \exp\left[\hat{\beta}_{SMK} \pm 1.96\sqrt{\text{var}(\hat{\beta}_{SMK})}\right] &= \exp(1.4361 \pm 1.96 \times .3167) \\ &= (e^{0.8154}, e^{2.0568}) = (2.26, 7.82). \end{aligned}$$

The above statistical information gives a meaningfully significant point estimate of 4.20, which is statistically significant. The confidence interval is quite wide, ranging between 2 and 8, but the lower limit of 2 suggests that there is real effect of smoking in this data set. All of these results depend on the no interaction model being the correct model, however.

Chapter 9**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. Odds = $\exp[\alpha_2 + \beta_{21}(40) + \beta_{22}(1) + \beta_{23}(0) + \beta_{24}(\text{HPT})] = \exp[\alpha_2 + 40\beta_{21} + \beta_{22} + (\text{HPT})\beta_{24}]$
7. OR = $\exp(\beta_{12})$
8. OR = $\exp[(50-20)\beta_{21}] = \exp(30\beta_{21})$
9. $H_0: \beta_{13} = \beta_{23} = \beta_{33} = \beta_{14} = \beta_{24} = \beta_{34} = 0$
Test statistic: $-2 \log \text{likelihood of the model without the smoking and hypertension terms (i.e., the reduced model), minus } -2 \log \text{likelihood of the model containing the smoking and hypertension terms (i.e., the full model from Question 6).}$

Under the null hypothesis the test statistic follows an approximate chi-square distribution with six degrees of freedom.

10. $\ln\left[\frac{P(D = g \mid \mathbf{X})}{P(D = 0 \mid \mathbf{X})}\right] = [\alpha_g + \beta_{g1}\text{AGE} + \beta_{g2}\text{GENDER} + \beta_{g3}\text{SMOKE} + \beta_{g4}\text{HPT} + \beta_{g5}(\text{AGE} \times \text{SMOKE}) + \beta_{g6}(\text{GENDER} \times \text{SMOKE})]$
where $g = 1, 2, 3$

Six additional parameters are added to the interaction model ($\beta_{15}, \beta_{25}, \beta_{35}, \beta_{16}, \beta_{26}, \beta_{36}$).

Chapter 10**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 11**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 12

1. Model 1: logit $P(D=1|\mathbf{X}) = \beta_0 + \beta_1 RX + \beta_2 SEQUENCE + \beta_3 RX*SEQUENCE$
 Model 2: logit $P(D=1|\mathbf{X}) = \beta_0 + \beta_1 RX + \beta_2 SEQUENCE$
 Model 3: logit $P(D=1|\mathbf{X}) = \beta_0 + \beta_1 RX$
2. Estimated OR (where SEQUENCE = 0) = $\exp(0.4184) = 1.52$
 95% CI: $\exp[0.4184 \pm 1.96(0.5885)] = (0.48, 4.82)$
 Estimated OR (where SEQUENCE = 1) = $\exp(0.4184 - 0.2136) = 1.23$
 $95\% \text{ CI: } \exp[(0.4184 - 0.2136) \pm 1.96 \sqrt{0.3463 + 0.6388 - 2(0.3463)}] = (0.43, 3.54)$
Note: var($\hat{\beta}_1 + \hat{\beta}_3$) = var($\hat{\beta}_1$) + var($\hat{\beta}_3$) + 2 cov($\hat{\beta}_1, \hat{\beta}_3$). See Chapter 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: $z = (0.2136 / 0.7993) = 0.2672$; $P\text{-value} = 0.79$
 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-squared with one degree of freedom); P -value = 0.79

Conclusion: do not reject H_0 .

6. \widehat{OR} (from Model 2): $\exp(0.3104) = 1.36$;

\widehat{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 13

True-False Questions:

1. F; a marginal model does not include a subject specific effect
2. T
3. T
4. T
5. T
6. $\text{logit } \mu_i = \beta_0 + \beta_1 RX_i + \beta_2 \text{SEQUENCE}_i + \beta_3 RX_i \times \text{SEQ}_i + b_{0i}$
7. \widehat{OR} (where SEQUENCE = 0) = $\exp(0.4707) = 1.60$
 \widehat{OR} (where SEQUENCE = 1) = $\exp(0.4707 - 0.2371) = 1.26$
8. $\widehat{OR} = \exp(0.3553) = 1.43$
 $95\% \text{ CI : } \exp[0.3553 \pm 1.96(0.4565)] = (0.58, 3.49)$
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.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W., *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, MA, 1975.

Breslow, N.R. and Clayton, D.G., Approximate inference in generalized linear mixed models, *J. Am. Statist. 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.

*Sources for practice exercises or test questions presented at the end of several chapters.

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, *Statist. Med.* 20:1461–1467, 2001.

Carey, V., Zeger, S.L., and Diggle, P., Modelling multivariate binary data with alternating logistic regressions, *Biometrika* 80(3):517–526, 1991.

*Donovan, J.W., MacLennan, R., and Adena, M., Vietnam service and the risk of congenital anomalies. A case-control study. *Med. J. Austr.* 140(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., Autsin, 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.

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., Muller, K.A., and Nizam, A., *Applied Regression Analysis and Other Multivariate Methods*, 3rd ed., Duxbury Press, Pacific Grove, CA, 1998.

Liang, K.Y. and Zeger, S.L., Longitudinal data analysis using generalized linear models, *Biometrika* 73:13–22, 1986.

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. Statist. Soc. B.* 42(2):109–142, 1980.

McCullagh, P. and Nelder, J.A., *Generalized Linear Models*, 2nd ed., Chapman & Hall, London, 1989.

*McLaws, N., Irwig, L.M., Mock, P., Berry, G., and Gold, J., Predictors of surgical wound infection in Australia: A national study, *Med. J. Austr.* 149:591–595, 1988.

Prentice, R.L. and Pyke, R., Logistic disease incidence models and case-control studies, *Biometrics* 32(3):599–606.

SAS Institute, *SAS/STAT User's Guide, Version 8.0*, SAS Institute, Inc., Cary, NC, 2000.

Wolfinger, R. and O'Connell, M., Generalized linear models: A pseudo-likelihood approach, *J. Statist. Comput. Simul.* 48:233–243, 1993.

Index

A

- Additive interaction, 50
- Adjusted odds ratio, 26, 27, 77
- ALR. *See* Alternating logistic regressions
- Alternating logistic regressions (ALR), 408, 409–413
- Aspirin-Heart Bypass Study, 389–393, 410–413
- Asymptotic properties, 354
- Autoregressive correlation structure, 350
- AR1, 336, 351, 383

B

- Background odds, 20

Backward elimination approach, 174, 175

Baseline odds, 20

“Best” model, guidelines for, 164

Biased estimates, 115

Binomial-based model, 110

Black/white Cancer Survival Study, 272, 439

Block diagonal matrix, 347

C

- Case-control study, 11–15, 230, 238–241
- intercept in, 114
- pair-matched, 63–64, 107–109, 233
- Category matching, 231

- Chi-square tests, 130
 Mantel-Haenszel, 234
- Chunk test, 195–196
- CLR. *See* Conditional logistic regression
- Coding
 for dichotomous variables, 77–82
 tests dependent on, 173
 tests independent of, 172–173
- Computer data sets, 437–440
- Conditional estimate, 241
- Conditional likelihood, 105–109, 113–114
- Conditional logistic regression, 236, 241, 413–417
 SAS and, 451–454
 SPSS and, 468–469
 Stata and, 481
- Confidence intervals, 119
 estimation of, 118, 136–137
 interactions and, 138–142
 large-sample formula, 119, 285, 311
 matching, 231
 narrower, 202
 odds ratios and, 137, 310–313
 one coefficient, 136–138
- Confounding
 assessment, 167–168, 192–220
 general procedure, 204
 interactions and, 203–211
 modeling strategy for, 192–220
 odds ratio and, 87–91, 209
 potential, 53, 54, 57, 62, 237
 precision and, 199
- Consistency, 354, 403
- Control variables, 52
- Controls, 230
- Correlation
 autoregressive, 336, 350–351, 383
 block diagonal matrix and, 347
 covariance and, 338–340
 defined, 338
 exchangeable, 349
 fixed, 353
- independent, 349
 matrix forms, 345–347
 stationary m-dependent, 352
 structures, 345–348
 types of, 349–354
 unstructured, 348, 352–353
See also specific procedures, parameters
- Covariance
 correlation and, 338–340
 defined, 338
 matrix forms, 115–116, 128–129, 345
- Cox proportional hazards regression, 165
- Cross-sectional studies, 11–15
- D**
- Data layout for correlated analysis, 337
- Data sets, computer. *See Computer data sets*
- Degrees of freedom, 287
- Dependence, and coding, 172–173
- Diagonal matrix, 346
- Dichotomous data, 5, 278, 340–341
 coding schemes for, 77–82
 GEE and, 459–464, 484–486
- Discriminant function analysis, 105
- Dispersion, 361
- Dummy variables, 61, 237
 interaction and, 243
- E**
- E, V, W model, 52–61
 general, 53–54
 logit form of, 76
 for several exposure variables, 85–87
- Effect modifiers, 62, 243
- Eligible subsets, 202
- Elimination, candidates for, 204
- Epidemiologic study framework, 8
- Estimated variance-covariance matrix, 115–116, 128–129

Evans County Heart Disease Study, 58, 59, 116, 142, 154, 177, 211–218, 438

Exchangeable sets, 245

Exchangeable correlation structure, 349

Exposure variables, 4, 43

 EVW model, 85–87

 odds ratio for, 87–91

 single, 42

Extra variation, 361

F

Fixed correlation structure, 353

Fixed effects, 414, 417

Follow-up studies, 11–15

Full model, 131

G

GEE. *See* Generalized estimating equations

General odds ratio formula, 84

Generalized linear mixed model (GLMM), 418–425

Generalized linear models (GLM), 341–343,

364

Generalizing estimating equations (GEE)

model, 330–403

 ALR model and, 407, 412

 Aspirin-Heart Bypass Study, 389–393

 asymptotic properties of, 354

 defined, 344–345

 dichotomous data and, 459–464, 484–486

 GLM and, 364

 Heartburn Relief Study, 393–395

 Infant Care Study, 380–388

 matrix notation for, 362–363

 parameters in, 364

 score-like equations and, 362–366

 statistical properties of, 354

GLMM. *See* Generalized linear mixed model

GLM. *See* Generalized linear models

Gold standard model, 200–203, 207

H

Hat symbol, 9

Heartburn Relief Study, 393–395

Hierarchical backward-elimination approach, 174, 175

Hierarchically well-formulated (HWF) model, 171–173, 244

Hierarchy principle, 175

 applying, 204

 product terms and, 180

 rationale for, 176

 retained variables, 175–176

HWF. *See* Hierarchically well-formulated model

Hypothesis testing, 9, 115, 118, 128–136

I

Identity matrix, 349

Independent correlation structure, 349

Independent of coding tests, 172–173

Independent variables, 4–5

Indicator variables, 61, 236–237

Infant care study, 380–388, 440

Inference, statistical, 115–119, 126–145, 279–282. *See also specific methods*

Influential observation, 169

Interactions

 additive, 50

 assessment of, 166–167, 179, 195–198

 coefficients of, 62

 confidence interval estimation with, 138–142

 confounding and, 203–211

 dummy variables and, 243

 likelihood ratio test, 146

 matching, 242–244

 modeling strategy for, 192–220

 multiplicative, 46–52

 no interaction, 48–49, 60, 199–203

 odds ratio for several exposure variables with, 87–91

Interactions (*continued*)
 precision and, 203–211
 product terms, 243
 variables for, 53
 Wald tests, 286

Intercept term, 82, 114

Interval estimation. *See* Confidence interval estimation

Interval variables, 79

Invariance, 305–306

Iterative methods, 111

J

Joint probability, 110, 112, 289

L

L. *See* Likelihood function

Large-sample formula, 119, 285, 311

Large versus small number of parameters debate, 106–108

Least squares (LS) estimation, 104

Likelihood function (L), 109–115

- for conditional method, 112, 113
- for ordinance model, 316–317
- for polytomous model, 288–290
- for unconditional method, 112

Likelihood ratio (LR) statistic, 130–134, 287, 357

- carrying out, 144

- defined, 118

- interaction terms, 146

Likelihood statistic, log, 130

Linear regression, 165

Link function, 343, 344

Log odds, logit and, 19–21

Logistic function, shape of, 6–7

Logistic model, 5–8

- application of, 9–11

- defined, 8

- follow-up study and, 14–15

- interaction, 46–52, 84

matching and, 61–64, 235–238
 multiplicative interaction, 46–52
 simple analysis, 43–46
 special cases of, 40–64
See also Logistic regression

Logistic regression

- ALR, 409–413

- basic features of, 4–7

- computing odds ratio in, 74–91

- conditional, 236, 241, 413–417, 451–454, 468–469, 481

- defined, 5

- introduction to, 2–31

- matching and, 230–242

- multiple standard, 317–319

- ordinal, 304, 319, 456–459, 471–473, 483

- polytomous, 272–291, 454–456, 469–471, 482

- statistical inferences for, 115–119, 126–145

- stratified analysis, 236

- unconditional, 441–451, 465–468, 475–480

- See also* Logistic model

Logit form, of EVW model, 76

Logit transformation, 16–22

- log odds and, 19–21

- logistic model and, 17

Log likelihood statistic, 130

LR statistic. *See* Likelihood ratio

LS. *See* Least squares estimation

M

Main effect variables, 27, 50

Mantel-Haenszel odds ratio (MOR), 234–235, 249

Marginal model, 423, 424

Matching, 114

- application of, 238–241

- basic features of, 230–232

- case-control studies, 230–247

- category matching, 231–232

- cohort studies, 247–249

- confidence intervals, 231

- exchangeability and, 245–246

- follow-up data, 247–251
 interaction and, 242–244
 logistic model and, 235–238
 major disadvantage, 232
 matching factors, 230
 ML estimation and, 239
 pooling, 245–247
 precision and, 231
 stratification, 232–235
 validity and, 232
 Mathematical models, 5
 Matrix algebra, 115–116, 128–129,
 345, 362
 Maximum likelihood (ML) methods,
 104–115
 numerical example, 143–149
 overview, 102–119
 statistical inferences using, 126–142
 unconditional versus conditional,
 105–109
 McNemar test, 234–235, 249
 Meaningful change in OR, concept of, 205
 MI dataset, 438
 Mixed logistic model (MLM), 418, 422
 ML. *See* Maximum likelihood methods
 MLM. *See* Mixed logistic model
 Modeling strategy
 confounding and, 192–220
 interaction and, 192–220
 example of, 177–181
 guidelines for, 162–182
 overview of, 164–169
 rationale for, 164–165
 Moderate samples, 119
 MOR. *See* Mantel-Haenszel odds ratio
 Multicollinearity, 168
 Multilevel outcomes, 270
 Multiple linear regression, 165
 Multiple standard logistic regressions, 291,
 317–319
 Multiple testing, 168
 Multiplicative interaction, 46–52
 Multivariable problem, 4–5
- N**
 No interaction model, 48–49, 60, 199–203
 Nominal exposure variable, 82–84
 Normal distribution, 104
 Nuisance parameters, 114
 Null hypotheses, 51, 280, 281
- O**
 Odds, 18–19
 Odds ratio (OR), 11–13
 adjusted, 26, 27, 77
 computation of, 25–26, 64, 74–91
 confidence limits, 137, 310–313
 confounders and, 87–91, 209
 correlation measure and, 409
 examples of, 22–23
 exchangeable, 411
 as fixed, 57, 79
 formula for, 22–25, 84
 invariance of, 305, 306
 logistic regression and, 74–91
 MOR and, 234–235
 risk ratio and, 15–16
 three categories, 275–279
 One-to-one matching, 231
 OR. *See* Odds ratio
 Ordinal logistic regression, 304–309
 SAS and, 456–459
 SPSS and, 471–473
 Stata and, 483
 Ordinal models, 304–310
 Ordinal variables, 79
- P**
 Pair matching, 108–109, 231–236,
 238–241
 Parameterizing, of model, 308
 Parameters, number of, 106
 Polytomous logistic regression, 272–275
 adding variables, 282–284
 extending, 282–287

Polytomous logistic regression (*continued*)
 likelihood function for, 288–290
 odds ratios from, 278
 ordinal model and, 310
 proportional odds model, 306–307
 SAS and, 454–456
 SPSS and, 469–471
 Stata and, 482
 Pooling, 245–247
 Potential confounders, 53, 54, 57, 62, 237
 Precision
 confounding and, 199
 consideration of, 167, 210–211
 gaining, 202
 interaction and, 203–211
 matching and, 231
 validity and, 167–168
 Predicted risk, 10
 Prediction, 165
 Probability, 43, 110
 Product terms, 28, 62, 117, 170
 hierarchy principle, 179–181, 185
 interaction and, 198, 237, 243
 Proportional odds model, 304–310
 alternate formulation of, 307
 polytomous model, 306–307

Q

Quasi-likelihood
 estimating equations, 360
 methods, 344

R

R-to-1 matching, 231
 Random effects, 417–423
 Rare disease assumption, 16
 Reduced model, 51, 131
 Referent group, 47, 230
 Retaining variables, 175–176
 Risk, 10, 43

Risk estimates, 13
 Risk odds ratio (ROR), 23–24
 estimated, 26
 general formula for, 44–45
 product formula for, 25
 Risk ratio (RR), 11–13, 15–16
 Robust conditions, 12
 ROR. *See* Risk odds ratio
 RR. *See* Risk ratio

S

Sample size, 119
 SAS software, 105, 308, 437, 441–464
 Scale factor, 360, 364–365
 Score equations, 359–361
 Score-like equations, 359–363
 Score statistic, 136
 Score test, 310–311, 318, 357–361
 Simple analysis, 43–46
 Single exposure variables, 42
 Small samples, 119
 Small versus large number of parameters
 debate, 106–108
 Software packages. *See* Computer data sets;
specific programs
 SPSS software, 105, 308, 437, 464–473
 Standard error, estimated, 136–137
 Standard logistic regression, 279
 Stata software, 105, 308, 437, 474–486
 stationary m-dependent correlation structure,
 352
 Statistical inference, 115–119,
 126–145
 Statistical tests for GEE, 357–358
 Stratification, 61, 114
 logistic regression, 236
 matching and, 232–235
 Study variables, 4
 Subject-specific effects, 418–423
 Subsets, eligible, 202
 Symmetric matrix, 346

T

- Threshold idea, 7
Time-dependent variables, 332
Time-independent variables, 332

U

- Unbiased estimates, 115
Unconditional estimate, 241
Unconditional logistic regression
 SAS and, 441–451
 SPSS and, 465–468
 Stata and, 475–480
Unconditional ML approach, 105–109
Unknown parameters, 8
Unstructured correlation structure, 348,
 352–353

V

- Validity, 165

matching and, 232

precision and, 167–168

Variable specification, 165, 169–171

Variables, retaining, 175–176

Variance-covariance matrix, 115–116,
 128–129

Variance estimators, 354–357

W

- Wald tests, 134–136
 carrying out, 144
 defined, 118
 GEE models and, 357
 ordinal model and, 313
 polytomous model and, 281, 286

Z

- Z statistic, 135–136