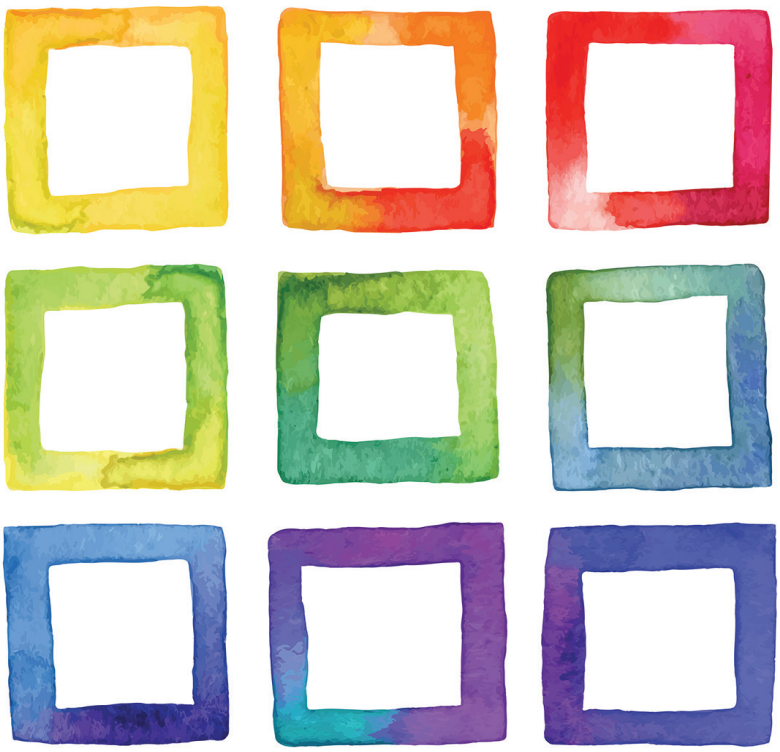
Wiley Series in Probability and Statistics

AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS

THIRD EDITION



ALAN AGRESTI

with website

WILEY

AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS

WILEY SERIES IN PROBABILITY AND STATISTICS
Established by *Walter A. Shewhart and Samuel S. Wilks*

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches. This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

A complete list of titles in this series can be found at
http://www.wiley.com/go/wsps

AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS

Third Edition

Alan Agresti
University of Florida, Florida, United States

WILEY

This third edition first published 2019

Edition History

(1e, 1996); John Wiley & Sons, Inc. (2e, 2007); John Wiley & Sons, Inc.

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data
Names: Agresti, Alan, author.
Title: An introduction to categorical data analysis / Alan Agresti.
Description: Third edition. | Hoboken, NJ : John Wiley & Sons, 2019. | Series: Wiley series in probability and statistics | Includes bibliographical references and index. |
Identifiers: LCCN 2018026887 (print) | LCCN 2018036674 (ebook) | ISBN 9781119405276 (Adobe PDF) | ISBN 9781119405283 (ePub) | ISBN 9781119405269 (hardcover)
Subjects: LCSH: Multivariate analysis.
Classification: LCC QA278 (ebook) | LCC QA278 .A355 2019 (print) | DDC 519.5/35–dc23
LC record available at https://lccn.loc.gov/2018026887

Cover Design: Wiley
Cover Image: © iStock.com/Anna_Zubkova

Set in 10/12.5pt Nimbus by Aptara Inc., New Delhi, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

In recent years, the use of specialized statistical methods for categorical data has increased dramatically, particularly for applications in the biomedical and social sciences. Partly this reflects the development during the past few decades of sophisticated methods for analyzing categorical data. It also reflects the increasing methodological sophistication of scientists and applied statisticians, most of whom now realize that it is unnecessary and often inappropriate to use methods for continuous data with categorical responses.

This third edition of the book is a substantial revision of the second edition. The most important change is showing how to conduct all the analyses using `R` software. As in the first two editions, the main focus is presenting the most important methods for analyzing categorical data. The book summarizes methods that have long played a prominent role, such as chi-squared tests, but gives special emphasis to modeling techniques, in particular to logistic regression.

The presentation in this book has a low technical level and does not require familiarity with advanced mathematics such as calculus or matrix algebra. Readers should possess a background that includes material from a two-semester statistical methods sequence for undergraduate or graduate nonstatistics majors. This background should include estimation and significance testing and exposure to regression modeling.

This book is designed for students taking an introductory course in categorical data analysis, but I also have written it for applied statisticians and practicing scientists involved in data analyses. I hope that the book will be helpful to analysts dealing with categorical response data in the social, behavioral, and biomedical sciences, as well as in public health, marketing, education, biological and agricultural sciences, and industrial quality control.

The basics of categorical data analysis are covered in Chapters 1 to 7. Chapter 2 surveys standard descriptive and inferential methods for contingency tables, such as odds ratios, tests

of independence, and conditional versus marginal associations. I feel that an understanding of methods is enhanced, however, by viewing them in the context of statistical models. Thus, the rest of the text focuses on the modeling of categorical responses. I prefer to teach categorical data methods by unifying their models with ordinary regression models. Chapter 3 does this under the umbrella of generalized linear models. That chapter introduces generalized linear models for binary data and count data. Chapters 4 and 5 discuss the most important such model for binary data, logistic regression. Chapter 6 introduces logistic regression models for multicategory responses, both nominal and ordinal. Chapter 7 discusses loglinear models for contingency tables and other types of count data.

I believe that logistic regression models deserve more attention than loglinear models, because applications more commonly focus on the relationship between a categorical response variable and some explanatory variables (which logistic regression models do) than on the association structure among several response variables (which loglinear models do). Thus, I have given main attention to logistic regression in these chapters and in later chapters that discuss extensions of this model.

Chapter 8 presents methods for matched-pairs data. Chapters 9 and 10 extend the matched-pairs methods to apply to clustered, correlated observations. Chapter 9 does this with marginal models, emphasizing the generalized estimating equations (GEE) approach, whereas Chapter 10 uses random effects to model more fully the dependence. Chapter 11 is a new chapter, presenting classification and smoothing methods. That chapter also introduces regularization methods that are increasingly important with the advent of data sets having large numbers of explanatory variables. Chapter 12 provides a historical perspective of the development of the methods. The text concludes with an appendix showing the use of `R`, SAS, Stata, and SPSS software for conducting nearly all methods presented in this book. Many of the chapters now also show how to use the Bayesian approach to conduct the analyses.

The material in Chapters 1 to 7 forms the heart of an introductory course in categorical data analysis. Sections that can be skipped if desired, to provide more time for other topics, include Sections 1.5, 2.5–2.7, 3.3 and 3.5, 5.4–5.6, 6.3–6.4, and 7.4–7.6. Instructors can choose sections from Chapters 8 to 12 to supplement the topics of primary importance. Sections and subsections labeled with an asterisk can be skipped for those wanting a briefer survey of the methods.

This book has lower technical level than my book *Categorical Data Analysis* (3rd edition, Wiley 2013). I hope that it will appeal to readers who prefer a more applied focus than that book provides. For instance, this book does not attempt to derive likelihood equations, prove asymptotic distributions, or cite current research work.

Most methods for categorical data analysis require extensive computations. For the most part, I have avoided details about complex calculations, feeling that statistical software should relieve this drudgery. The text shows how to use `R` to obtain all the analyses presented. The Appendix discusses the use of SAS, Stata, and SPSS. The full data sets analyzed in the book are available at the text website `www.stat.ufl.edu/~aa/cat/data`. That website also lists typos and errors of which I have become aware since publication. The data files are also available at `https://github.com/alanagresti/categorical-data`.

Brief solutions to odd-numbered exercises appear at the end of the text. An instructor's manual will be included on the companion website for this edition: `www.wiley.com/go/Agresti/CDA_3e`. The aforementioned data sets will also be available on the companion website. Additional exercises are available there and at `www.stat.ufl.edu/`

`~aa/cat/Extra_Exercises`, some taken from the 2nd edition to create space for new material in this edition and some being slightly more technical.

I owe very special thanks to Brian Marx for his many suggestions about the text over the past twenty years. He has been incredibly generous with his time in providing feedback based on teaching courses based on the book. I also thank those individuals who commented on parts of the manuscript or who made suggestions about examples or material to cover or provided other help such as noticing errors. Travis Gerke, Anna Gottard, and Keramat Nourijelyani gave me several helpful comments. Thanks also to Alessandra Brazzale, Debora Giovannelli, David Groggel, Stacey Handcock, Maria Kateri, Bernhard Klingenberg, Ioannis Kosmidis, Mohammad Mansournia, Trevelyan McKinley, Changsoon Park, Tom Piazza, Brett Presnell, Ori Rosen, Ralph Scherer, Claudia Tarantola, Anestis Touloumis, Thomas Yee, Jin Wang, and Sherry Wang. I also owe thanks to those who helped with the first two editions, especially Patricia Altham, James Booth, Jane Brockmann, Brian Caffo, Brent Coull, Al DeMaris, Anna Gottard, Harry Khamis, Svend Kreiner, Carla Rampichini, Stephen Stigler, and Larry Winner. Thanks to those who helped with material for my more advanced text (*Categorical Data Analysis*) that I extracted here, especially Bernhard Klingenberg, Yongyi Min, and Brian Caffo. Many thanks also to the staff at Wiley for their usual high-quality help.

A truly special by-product for me of writing books about categorical data analysis has been invitations to teach short courses based on them and spend research visits at many institutions around the world. With grateful thanks I dedicate this book to my hosts over the years. In particular, I thank my hosts in Italy (Adelchi Azzalini, Elena Beccalli, Rino Bellocco, Matilde Bini, Giovanna Boccuzzo, Alessandra Brazzale, Silvia Cagnone, Paula Cerchiello, Andrea Cerioli, Monica Chiogna, Guido Consonni, Adriano Decarli, Mauro Gasparini, Alessandra Giovagnoli, Sabrina Giordano, Paolo Giudici, Anna Gottard, Alessandra Guglielmi, Maria Iannario, Gianfranco Lovison, Claudio Lupi, Monia Lupparelli, Maura Mezzetti, Antonietta Mira, Roberta Paroli, Domenico Piccolo, Irene Poli, Alessandra Salvan, Nicola Sartori, Bruno Scarpa, Elena Stanghellini, Claudia Tarantola, Cristiano Varin, Roberta Varriale, Laura Ventura, Diego Zappa), the UK (Phil Brown, Bianca De Stavola, Brian Francis, Byron Jones, Gillian Lancaster, Irini Moustaki, Chris Skinner, Briony Teather), Austria (Regina Dittrich, Gilg Seeber, Helga Wagner), Belgium (Hermann Callaert, Geert Molenberghs), France (Antoine De Falguerolles, Jean-Yves Mary, Agnes Rogel), Germany (Maria Kateri, Gerhard Tutz), Greece (Maria Kateri, Ioannis Ntzoufras), the Netherlands (Ivo Molenaar, Marijte van Duijn, Peter van der Heijden), Norway (Petter Laake), Portugal (Francisco Carvalho, Adelaide Freitas, Pedro Oliveira, Carlos Daniel Paulino), Slovenia (Janez Stare), Spain (Elias Moreno), Sweden (Juni Palmgren, Elisabeth Svensson, Dietrich van Rosen), Switzerland (Anthony Davison, Paul Embrechts), Brazil (Clarice Demetrio, Bent Jörgensen, Francisco Louzada, Denise Santos), Chile (Guido Del Pino), Colombia (Marta Lucia Corrales Bossio, Leonardo Trujillo), Turkey (Aylin Alin), Mexico (Guillermina Eslava), Australia (Chris Lloyd), China (I-Ming Liu, Chongqi Zhang), Japan (Ritei Shibata), and New Zealand (Nye John, I-Ming Liu). Finally, thanks to my wife, Jacki Levine, for putting up with my travel schedule in these visits around the world!

ALAN AGRESTI

Gainesville, Florida and Brookline, Massachusetts
March 2018

ABOUT THE COMPANION WEBSITE

This book comes with a companion website of other material, including all data sets analyzed in the book and some extra exercises.

www.wiley.com/go/Agresti/CDA_3e 

CHAPTER 1

INTRODUCTION

From helping to assess the value of new medical treatments to evaluating the factors that affect our opinions on controversial issues, scientists today are finding myriad uses for categorical data analyses. It is primarily for these scientists and their collaborating statisticians – as well as those training to perform these roles – that this book was written.

This first chapter reviews the most important probability distributions for categorical data: the *binomial* and *multinomial* distributions. It also introduces *maximum likelihood*, the most popular method for using data to estimate parameters. We use this type of estimate and a related *likelihood function* to conduct statistical inference. We also introduce the *Bayesian* approach to statistical inference, which utilizes probability distributions for the parameters as well as for the data. We begin by describing the major types of categorical data.

1.1 CATEGORICAL RESPONSE DATA

A *categorical* variable has a measurement scale consisting of a set of categories. For example, political ideology might be measured as liberal, moderate, or conservative; choice of accommodation might use categories house, condominium, and apartment; a diagnostic test to detect e-mail spam might classify an incoming e-mail message as spam or legitimate. Categorical variables are often referred to as *qualitative*, to distinguish them from *quantitative* variables, which take numerical values, such as age, income, and number of children in a family.

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

Categorical variables are pervasive in the social sciences for measuring attitudes and opinions, with categories such as (agree, disagree), (yes, no), and (favor, oppose, undecided). They also occur frequently in the health sciences, for measuring responses such as whether a medical treatment is successful (yes, no), mammogram-based breast diagnosis (normal, benign, probably benign, suspicious, malignant with cancer), and stage of a disease (initial, intermediate, advanced). Categorical variables are common for service-quality ratings of any company or organization that has customers (e.g., with categories excellent, good, fair, poor). In fact, categorical variables occur frequently in most disciplines. Other examples include the behavioral sciences (e.g., diagnosis of type of mental illness, with categories schizophrenia, depression, neurosis), ecology (e.g., primary land use in satellite image, with categories woodland, swamp, grassland, agriculture, urban), education (e.g., student responses to an exam question, with categories correct, incorrect), and marketing (e.g., consumer cell-phone preference, with categories Samsung, Apple, Nokia, LG, Other). They even occur in highly quantitative fields such as the engineering sciences and industrial quality control, when items are classified according to whether or not they conform to certain standards.

1.1.1 Response Variable and Explanatory Variables

Most statistical analyses distinguish between a *response* variable and *explanatory* variables. For instance, ordinary regression models describe how the mean of a quantitative response variable, such as annual income, changes according to levels of explanatory variables, such as number of years of education and number of years of job experience. The response variable is sometimes called the *dependent variable* and the explanatory variable is sometimes called the *independent variable*. When we want to emphasize that the response variable is a random variable, such as in a probability statement, we use upper-case notation for it (e.g., Y). We use lower-case notation to refer to a particular value (e.g., $y = 0$).

This text presents statistical models that relate a categorical response variable to explanatory variables that can be categorical or quantitative. For example, a study might analyze how opinion about whether same-sex marriage should be legal (yes or no) is associated with explanatory variables such as number of years of education, annual income, political party affiliation, religious affiliation, age, gender, and race.

1.1.2 Binary–Nominal–Ordinal Scale Distinction

Many categorical variables have only two categories, such as (yes, no) for possessing health insurance or (favor, oppose) for legalization of marijuana. Such variables are called *binary variables*.

When a categorical variable has more then two categories, we distinguish between two types of categorical scales. Categorical variables having *unordered* scales are called *nominal* variables. Examples are religious affiliation (categories Christian, Jewish, Muslim, Buddhist, Hindu, none, other), primary mode of transportation to work (automobile, bicycle, bus, subway, walk), and favorite type of music (classical, country, folk, jazz, pop, rock). Variables having naturally *ordered* categories are called *ordinal* variables. Examples are perceived happiness (not too happy, pretty happy, very happy), frequency of feeling anxiety (never, occasionally, often, always), and headache pain (none, slight, moderate, severe).

A variable's measurement scale determines which statistical methods are appropriate. For nominal variables, the order of listing the categories is arbitrary, so methods designed for them give the same results no matter what order is used. Methods designed for ordinal variables utilize the category ordering.

1.1.3 Organization of this Book

Chapters 1 and 2 describe basic non model-based methods of categorical data analysis. These include analyses of proportions and of association between categorical variables.

Chapters 3 to 7 introduce models for categorical response variables. These models resemble regression models for quantitative response variables. In fact, Chapter 3 shows they are special cases of a class of *generalized linear models* that also contains the ordinary normal-distribution-based regression models. *Logistic regression* models, which apply to binary response variables, are the focus of Chapters 4 and 5. Chapter 6 extends logistic regression to multicategory responses, both nominal and ordinal. Chapter 7 introduces *loglinear* models, which analyze associations among multiple categorical response variables.

The methods in Chapters 1 to 7 assume that observations are independent. Chapters 8 to 10 introduce logistic regression models for observations that are correlated, such as for matched pairs or for repeated measurement of individuals in longitudinal studies. Chapter 11 introduces some advanced methods, including ways of classifying and clustering observations into categories and ways of dealing with data sets having huge numbers of variables. The book concludes (Chapter 12) with a historical overview of the development of categorical data methods.

Statistical software packages can implement methods for categorical data analysis. We illustrate throughout the text for the free software `R`. The Appendix discusses the use of SAS, Stata, and SPSS. A companion website for the book, `www.stat.ufl.edu/~aa/cat`, has additional information, including complete data sets for the examples. The data files are also available at `https://github.com/alanagresti/categorical-data`.

1.2 PROBABILITY DISTRIBUTIONS FOR CATEGORICAL DATA

Parametric inferential statistical analyses require an assumption about the probability distribution of the response variable. For regression models for quantitative variables, the normal distribution plays a central role. This section presents the key probability distributions for categorical variables: the *binomial* and *multinomial* distributions.

1.2.1 Binomial Distribution

When the response variable is binary, we refer to the two outcome categories as *success* and *failure*. These labels are generic and the *success* outcome need not be a preferred result.

Many applications refer to a fixed number n of independent and identical trials with two possible outcomes for each. *Identical trials* means that the probability of success is the same for each trial. *Independent trials* means the response outcomes are independent random variables. In particular, the outcome of one trial does not affect the outcome of another. These are often called *Bernoulli trials*. Let π denote the probability of success for

each trial. Let Y denote the number of successes out of the n trials. Under the assumption of n independent, identical trials, Y has the *binomial distribution* with index n and parameter π. The probability of a particular outcome y for Y equals

$$P(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}, \quad y = 0, 1, 2, \ldots, n. \tag{1.1}$$

To illustrate, suppose a quiz has ten multiple-choice questions, with five possible answers for each. A student who is completely unprepared randomly guesses the answer for each question. Let Y denote the number of correct responses. For each question, the probability of a correct response is 0.20, so $\pi = 0.20$ with $n = 10$. The probability of $y = 0$ correct responses, and hence $n - y = 10$ incorrect ones, equals

$$P(0) = \frac{10!}{0!10!}(0.20)^0(0.80)^{10} = (0.80)^{10} = 0.107.$$

The probability of 1 correct response equals

$$P(1) = \frac{10!}{1!9!}(0.20)^1(0.80)^9 = 10(0.20)(0.80)^9 = 0.268.$$

Table 1.1 shows the binomial distribution for all the possible values, $y = 0, 1, 2, \ldots, 10$. For contrast, it also shows the binomial distributions when $\pi = 0.50$ and when $\pi = 0.80$.

Table 1.1 Binomial distributions with $n = 10$ and $\pi = 0.20$, 0.50, and 0.80. The binomial distribution is symmetric when $\pi = 0.50$.

y	$P(y)$ when $\pi = 0.20$ ($\mu = 2.0$, $\sigma = 1.26$)	$P(y)$ when $\pi = 0.50$ ($\mu = 5.0$, $\sigma = 1.58$)	$P(y)$ when $\pi = 0.80$ ($\mu = 8.0$, $\sigma = 1.26$)
0	0.107	0.001	0.000
1	0.268	0.010	0.000
2	0.302	0.044	0.000
3	0.201	0.117	0.001
4	0.088	0.205	0.005
5	0.027	0.246	0.027
6	0.005	0.205	0.088
7	0.001	0.117	0.201
8	0.000	0.044	0.302
9	0.000	0.010	0.268
10	0.000	0.001	0.107

The binomial distribution for n trials with parameter π has mean and standard deviation

$$E(Y) = \mu = n\pi, \quad \sigma = \sqrt{n\pi(1-\pi)}.$$

The binomial distribution with $\pi = 0.20$ in Table 1.1 has $\mu = 10(0.20) = 2.0$. The standard deviation is $\sigma = \sqrt{10(0.20)(0.80)} = 1.26$, which σ also equals when $\pi = 0.80$.

The binomial distribution is symmetric when $\pi = 0.50$. For fixed n, it becomes more bell-shaped as π gets closer to 0.50. For fixed π, it becomes more bell-shaped as n increases.

When n is large, it can be approximated by a normal distribution with $\mu = n\pi$ and $\sigma = \sqrt{n\pi(1-\pi)}$. A guideline[1] is that the expected number of outcomes of the two types, $n\pi$ and $n(1-\pi)$, should both be at least about 5. For $\pi = 0.50$ this requires only $n \geq 10$, whereas $\pi = 0.10$ (or $\pi = 0.90$) requires $n \geq 50$. When π gets nearer to 0 or 1, larger samples are needed before a symmetric, bell shape occurs.

1.2.2 Multinomial Distribution

Nominal and ordinal response variables have more than two possible outcomes. When the observations are independent with the same category probabilities for each, the probability distribution of counts in the outcome categories is the *multinomial*.

Let c denote the number of outcome categories. We denote their probabilities by $(\pi_1, \pi_2, \ldots, \pi_c)$, where $\sum_j \pi_j = 1$. For n independent observations, the multinomial probability that y_1 fall in category 1, y_2 fall in category 2, ... , y_c fall in category c, where $\sum_j y_j = n$, equals

$$P(y_1, y_2, ..., y_c) = \left(\frac{n!}{y_1!y_2!\cdots y_c!}\right)\pi_1^{y_1}\pi_2^{y_2}\cdots\pi_c^{y_c}.$$

The binomial distribution is the special case with $c = 2$ categories. We will not need to use this formula, because our focus is on inference methods that use *sampling distributions* of statistics computed from the multinomial counts, and those sampling distributions are approximately *normal* or *chi-squared*.

1.3 STATISTICAL INFERENCE FOR A PROPORTION

In practice, the parameter values for binomial and multinomial distributions are unknown. Using sample data, we estimate the parameters. This section introduces the *maximum likelihood* estimation method and illustrates it for the binomial parameter.

1.3.1 Likelihood Function and Maximum Likelihood Estimation

The parametric approach to statistical modeling assumes a family of probability distributions for the response variable, indexed by an unknown parameter. For a particular family, we can substitute the observed data into the formula for the probability function and then view how that probability depends on the unknown parameter value. For example, in $n = 10$ trials, suppose a binomial count equals $y = 0$. From the binomial formula (1.1) with parameter π, the probability of this outcome equals

$$P(0) = \frac{10!}{0!10!}\pi^0(1-\pi)^{10} = (1-\pi)^{10}.$$

This probability is defined for all the potential values of π between 0 and 1.

[1] You can explore this with the binomial distribution applet at `www.artofstat.com/webapps.html`.

The probability of the observed data, expressed as a function of the parameter, is called the *likelihood function*. With $y = 0$ successes in $n = 10$ trials, the binomial likelihood function is $\ell(\pi) = (1 - \pi)^{10}$, for $0 \leq \pi \leq 1$. If $\pi = 0.40$, for example, the probability that $y = 0$ is $\ell(0.40) = (1 - 0.40)^{10} = 0.006$. Likewise, if $\pi = 0.20$ then $\ell(0.20) = (1 - 0.20)^{10} = 0.107$, and if $\pi = 0.0$ then $\ell(0.0) = (1 - 0.0)^{10} = 1.0$. Figure 1.1 plots this likelihood function for all π values between 0 and 1.

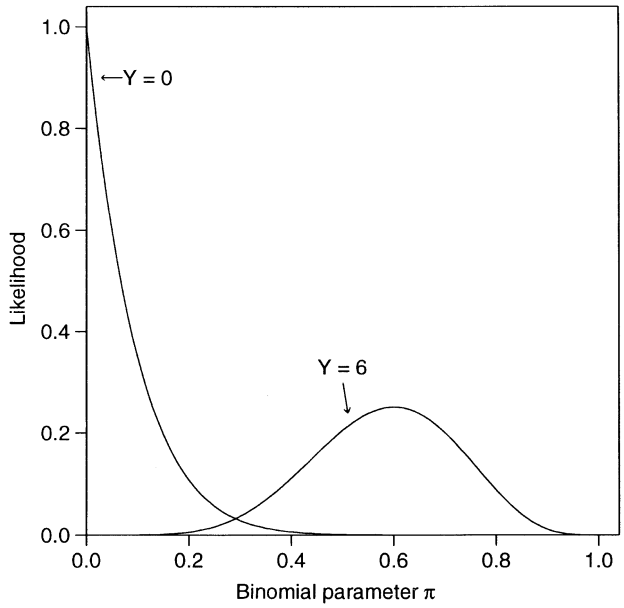


Figure 1.1 Binomial likelihood functions for $y = 0$ successes and for $y = 6$ successes in $n = 10$ trials.

The *maximum likelihood estimate* of a parameter is the parameter value at which the likelihood function takes its maximum. That is, it is the parameter value for which the probability of the observed data takes its greatest value. Figure 1.1 shows that the likelihood function $\ell(\pi) = (1 - \pi)^{10}$ has its maximum at $\pi = 0.0$. Therefore, when $n = 10$ trials have $y = 0$ successes, the maximum likelihood estimate of π equals 0.0. This means that the result $y = 0$ in $n = 10$ trials is more likely to occur when $\pi = 0.00$ than when π equals any other value.

We use the abbreviation *ML* to symbolize *maximum likelihood*. The ML estimate is often denoted by the parameter symbol with a ^ (a *hat*) over it. We denote the ML estimate of the binomial parameter π by $\hat{\pi}$, called *pi-hat*. In general, for the binomial outcome of y successes in n trials, the maximum likelihood estimate of π is $\hat{\pi} = y/n$. This is the sample proportion of successes for the n trials. If we observe $y = 6$ successes in $n = 10$ trials, then the maximum likelihood estimate of π is $\hat{\pi} = 6/10 = 0.60$. Figure 1.1 also plots the likelihood function when $n = 10$ with $y = 6$, which from formula (1.1) equals $\ell(\pi) = [10!/(6!4!)]\pi^6(1 - \pi)^4$. The maximum value occurs at $\hat{\pi} = 0.60$. The result $y = 6$ in $n = 10$ trials is more likely to occur when $\pi = 0.60$ than when π equals any other value.

If we denote each success by a 1 and each failure by a 0, then the sample proportion equals the sample mean of the data. For instance, for 4 failures followed by 6 successes in 10 trials, the data are $(0, 0, 0, 0, 1, 1, 1, 1, 1, 1)$ and the sample mean is

$$\hat{\pi} = (0 + 0 + 0 + 0 + 1 + 1 + 1 + 1 + 1 + 1)/10 = 0.60.$$

Thus, results that apply to sample means with random sampling apply also to sample proportions. These include the *Central Limit Theorem*, which states that the sampling distribution of the sample proportion $\hat{\pi}$ is approximately normal for large n, and the *Law of Large Numbers*, which states that $\hat{\pi}$ converges to the population proportion π as n increases.

Before we observe the data, the value of the ML estimate is unknown. The estimate is then a random variable having some sampling distribution. We refer to it as an *estimator* and its value for observed data as an *estimate*. Estimators based on the method of maximum likelihood are popular because they have good large-sample behavior. Sampling distributions of ML estimators are typically approximately normal and no other "good" estimator has a smaller standard error.

1.3.2 Significance Test About a Binomial Parameter

For the binomial distribution, we now use the ML estimator in statistical inference for the parameter π. The ML estimator $\hat{\pi}$ is the sample proportion. Its sampling distribution has mean and standard error

$$E(\hat{\pi}) = \pi, \;\; \sigma(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

Consider the null hypothesis H_0: $\pi = \pi_0$ that the parameter equals some fixed value, π_0, such as 0.50. When H_0 is true, the standard error of $\hat{\pi}$ is $SE_0 = \sqrt{\pi_0(1-\pi_0)/n}$, which we refer to as the *null standard error*. The test statistic

$$z = \frac{\hat{\pi} - \pi_0}{SE_0} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \tag{1.2}$$

divides the difference between the sample proportion $\hat{\pi}$ and the null hypothesis value π_0 by the null standard error. The z test statistic measures the number of standard errors that $\hat{\pi}$ falls from the H_0 value. For large samples, the null sampling distribution of z is the standard normal, which has mean $= 0$ and standard deviation $= 1$.

1.3.3 Example: Surveyed Opinions About Legalized Abortion

Do a majority, or minority, of adults in the United States believe that a pregnant woman should be able to obtain an abortion? Let π denote the proportion of the American adult population that responds *yes* when asked, "Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she wants it for any reason." We test H_0: $\pi = 0.50$ against the two-sided alternative hypothesis, H_a: $\pi \neq 0.50$.

This item was one of many about legalized abortion included in the 2016 General Social Survey (GSS). This survey, conducted every other year by the National Opinion Research Center (NORC) at the University of Chicago, asks a sample of adult Americans their opinions about a wide variety of issues.[2] The GSS is a multi-stage sample, but it has characteristics similar to a simple random sample. Of 1810 respondents to this item in 2016, 837 replied *yes* and 973 replied *no*. The sample proportion of *yes* responses was $\hat{\pi} = 837/1810 = 0.4624$.

[2] You can view responses to surveys since 1972 at `sda.berkeley.edu/archive.htm`.

The test statistic for H_0: $\pi = 0.50$ is

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.4624 - 0.50}{\sqrt{\frac{0.50(0.50)}{1810}}} = -3.20.$$

The two-sided P-value is the probability that the absolute value of a standard normal variate exceeds 3.20, which is $P = 0.0014$. The evidence is very strong that, in 2016, $\pi < 0.50$, that is, that fewer than half of Americans favored unrestricted legal abortion. In some other situations, such as when the mother's health was endangered, an overwhelming majority favored legalized abortion. Responses depended strongly on the question wording.

1.3.4 Confidence Intervals for a Binomial Parameter

A significance test merely indicates whether a particular value for a parameter (such as 0.50) is plausible. We learn more by constructing a confidence interval to determine the range of plausible values. Let $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$ denote the estimated standard error of $\hat{\pi}$. This formula obtains SE by substituting the ML estimate $\hat{\pi}$ for the unknown parameter π in $\sigma(\hat{\pi}) = \sqrt{\pi(1-\pi)/n}$. One way to form a $100(1-\alpha)\%$ confidence interval for π uses the formula

$$\hat{\pi} \pm z_{\alpha/2}(SE), \text{ with } SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}, \tag{1.3}$$

where $z_{\alpha/2}$ denotes the standard normal percentile having right-tail probability equal to $\alpha/2$; for example, for 95% confidence, $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$.

For the opinion about legalized abortion example just discussed, $\hat{\pi} = 0.462$ for $n = 1810$ observations. The 95% confidence interval equals

$$0.462 \pm 1.96\sqrt{0.462(0.538)/1810}, \text{ which is } 0.462 \pm 0.023, \text{ or } (0.439, 0.485).$$

We can be 95% confident that the population proportion of Americans in 2016 who favored unrestricted legalized abortion is between 0.439 and 0.485.

The significance test and confidence interval for π, as well as other confidence intervals presented next, are readily available in software and at web sites.[3]

1.3.5 Better Confidence Intervals for a Binomial Proportion *

Formula (1.3) is simple. When π is near 0 or near 1, however, it performs poorly unless n is very large. Its *actual* coverage probability, that is, the probability that the method produces an interval that captures the true parameter value, may be much less than the nominal value (such as 0.95).

A better way to construct confidence intervals uses a duality with significance tests. The confidence interval consists of all H_0 values π_0 that are judged plausible in the z test of Section 1.3.2. A 95% confidence interval contains all values π_0 for which the two-sided P-value exceeds 0.05. That is, it contains all values that are *not rejected* at the 0.05

[3] For instance, see `https://istats.shinyapps.io/Inference_prop`. The confidence interval (1.3) is the *Wald* type listed in the menu.

significance level. These are the H_0 values for π_0 that have test statistic z less than 1.96 in absolute value. This alternative method, called the *score confidence interval*, has the advantage that we do not need to estimate π in the standard error, because the standard error $SE_0 = \sqrt{\pi_0(1-\pi_0)/n}$ in the test statistic uses the null value π_0.

To illustrate, suppose that a clinical trial to evaluate a new treatment has 9 successes in the first 10 trials. For a sample proportion of $\hat{\pi} = 0.90$ based on $n = 10$, the value $\pi_0 = 0.596$ for the H_0 parameter value yields the test statistic value

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.90 - 0.596}{\sqrt{\frac{0.596(0.404)}{10}}} = 1.96$$

and a two-sided P-value of $P = 0.05$. The value $\pi_0 = 0.982$ yields

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.90 - 0.982}{\sqrt{\frac{0.982(0.018)}{10}}} = -1.96$$

and also a two-sided P-value of $P = 0.05$. All π_0 values between 0.596 and 0.982 have $|z| < 1.96$ and P-value > 0.05. Therefore, the 95% score confidence interval for π is (0.596, 0.982). For particular values of $\hat{\pi}$ and n, the π_0 values that have test statistic value $z = \pm 1.96$ are the solutions to the equation

$$\frac{|\,\hat{\pi} - \pi_0\,|}{\sqrt{\pi_0(1-\pi_0)/n}} = 1.96$$

for π_0. We will not deal here with how to solve this equation, as this confidence interval is readily available in software and at web sites.[4]

The simple formula (1.3) using estimated standard error fails spectacularly when $\hat{\pi} = 0$ or when $\hat{\pi} = 1$, regardless of how large n is. To illustrate, suppose the clinical trial had 10 successes in the 10 trials. Then, $\hat{\pi} = 10/10 = 1.0$ and $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{1.0(0.0)/10} = 0$, so the 95% confidence interval $1.0 \pm 1.96(SE)$ is 1.0 ± 0.0. This interval (1.0, 1.0) is completely unrealistic. When a sample estimate is at or near the boundary of the parameter space, having that estimate in the middle of the confidence interval results in poor performance of the method. By contrast, the 95% score confidence interval based on the corresponding significance test with null standard error SE_0 is (0.72, 1.0).

The score confidence interval itself has actual coverage probability a bit too small when π is very close to 0 or 1. A simple alternative confidence interval approximates the score interval but is a bit wider and has better coverage probability when π is near 0 or 1. It uses the simple formula (1.3) with the estimated standard error after adding 2 to the number of successes and 2 to the number of failures (and thus 4 to n). With 10 successes in 10 trials, you apply formula (1.3) to 12 successes in 14 trials and get (0.68, 1.0). This simple method,[5] called the *Agresti–Coull confidence interval*, has adequate coverage probability for small n even when π is very close to 0 or 1.

[4] Such as the "Wilson score" option at `https://istats.shinyapps.io/Inference_prop`.

[5] More precisely, software and the website `https://istats.shinyapps.io/Inference_prop` adds $(z^2_{\alpha/2})/2$ to each count (e.g., $(1.96)^2/2 = 1.92$ for 95% confidence); the CI then performs well because it has the same midpoint as the score CI but is a bit wider.

1.4 STATISTICAL INFERENCE FOR DISCRETE DATA

In summary, two methods we have presented for constructing a confidence interval for a proportion (1) use $\hat{\pi} \pm z_{\alpha/2}(SE)$ with the estimated standard error, $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$, or (2) invert results of a significance test using test statistic $z = (\hat{\pi} - \pi_0)/SE_0$ with the null standard error, $SE_0 = \sqrt{\pi_0(1-\pi_0)/n}$. These methods apply two of the three standard ways of conducting statistical inference (confidence intervals and significance tests) about parameters. We present the methods in a more general context in this section and also introduce a third standard inference method that uses the likelihood function.

1.4.1 Wald, Likelihood-Ratio, and Score Tests

Let β denote an arbitrary parameter, such as a linear effect of an explanatory variable in a model. Consider a significance test of H_0: $\beta = \beta_0$, such as H_0: $\beta = 0$ for which $\beta_0 = 0$. The simplest test statistic exploits the large-sample normality of the ML estimator $\hat{\beta}$. Let SE denote the unrestricted standard error of $\hat{\beta}$, evaluated by substituting the ML estimate for the unknown parameter in the expression for the true standard error. (For example, for the binomial parameter π, $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$.) When H_0 is true, the test statistic

$$z = (\hat{\beta} - \beta_0)/SE$$

has approximately a standard normal distribution. Equivalently, z^2 has approximately a chi-squared distribution with $df = 1$. This type of statistic, which uses the standard error evaluated at the ML estimate, is called a *Wald statistic*. The z test using this test statistic, or the corresponding chi-squared test that uses z^2, is called a *Wald test*.[6]

We can refer z to the standard normal distribution to get one-sided or two-sided P-values. For the two-sided alternative H_a: $\beta \neq \beta_0$, the P-value is also the right-tail chi-squared probability with $df = 1$ above the observed value of z^2. That is, the two-tail probability beyond $\pm z$ for the standard normal distribution equals the right-tail probability above z^2 for the chi-squared distribution with $df = 1$. For example, the two-tail standard normal probability of 0.05 that falls below -1.96 and above 1.96 equals the right-tail chi-squared probability above $(1.96)^2 = 3.84$ when $df = 1$. With the software `R` and its functions `pnorm` and `pchisq` for *cumulative probabilities* (i.e., probabilities *below* fixed values) for normal and chi-squared distributions, we find (with comments added following the # symbol):

```
------------------------------------------------------------------------------
> 2*pnorm(-1.96) # 2(standard normal cumulative probability below -1.96)
[1] 0.0499958    # essentially equals 0.05
> pchisq(1.96^2, 1) # pchisq gives chi-squared cumulative probability
[1] 0.9500042       # here, cumul. prob. at (1.96)(1.96) = 3.84 when df=1
> 1 - pchisq(1.96^2, 1) # right-tail prob. above (1.96)(1.96) when df=1
[1] 0.0499958    # same as normal two-tail probability
> # can also get this by  pchisq(1.96^2, 1, lower.tail=FALSE)
------------------------------------------------------------------------------
```

You can also find chi-squared and normal tail probabilities with applets on the Internet.[7]

[6] Proposed by the statistician Abraham Wald in 1943.

[7] See, for example, the applets at `www.artofstat.com/webapps.html#Distributions`.

A second possible test is called the *score test*.[8] This test uses standard errors that are valid when H_0 is true, rather than estimated more generally. For example, the z test (1.2) for a binomial parameter that uses the null standard error $SE_0 = \sqrt{\pi_0(1-\pi_0)/n}$ of $\hat{\pi}$ is a score test. The z test that uses $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$ instead of SE_0 is the Wald test.

A third possible test of H_0: $\beta = \beta_0$ uses the likelihood function through the ratio of two of its values. For a single parameter β, these are (1) the value ℓ_0 when H_0 is true (so $\beta = \beta_0$), (2) the maximum ℓ_1 over all possible parameter values, which is the likelihood function calculated at the ML estimate $\hat{\beta}$. Then ℓ_1 is always at least as large as ℓ_0, because ℓ_1 refers to maximizing over the entire parameter space rather than just at β_0. The *likelihood-ratio* test statistic[9] equals

$$2\log(\ell_1/\ell_0).$$

The reason for taking the log transform and doubling is that it yields an approximate chi-squared sampling distribution. Under H_0: $\beta = \beta_0$, this test statistic has a large-sample chi-squared distribution with $df = 1$. The test statistic $2\log(\ell_1/\ell_0)$ is nonnegative, and the P-value is the chi-squared right-tail probability. Larger values of (ℓ_1/ℓ_0) yield larger values of $2\log(\ell_1/\ell_0)$ and smaller P-values and stronger evidence against H_0.

For ordinary regression models that assume a normal distribution for Y, the Wald, score, and likelihood-ratio tests provide identical test statistics and P-values. For parameters in other statistical models, they have similar behavior when the sample size n is large and H_0 is true. When n is small to moderate, the Wald test is the least reliable of the three tests. The likelihood-ratio inference and score-test based inference are better in terms of actual inferential error probabilities, coming closer to matching nominal levels.

For any of the three tests, the P-value that software reports is an approximation for the true P-value. This is because the normal (or chi-squared) sampling distribution used is a large-sample approximation for the actual sampling distribution. Thus, when you report a P-value, it is overly optimistic to use many decimal places. If you are lucky, the P-value approximation is good to the second decimal place. Therefore, for a P-value that software reports as 0.028374, it makes more sense to report it as 0.03 (or, at best, 0.028) rather than 0.028374. An exception is when the P-value is zero to many decimal places, in which case it is sensible to report it as $P < 0.001$ or $P < 0.0001$. A P-value merely summarizes the strength of evidence against H_0, and accuracy to two or three decimal places is sufficient for this purpose.

Each significance test method has a corresponding confidence interval. The 95% confidence interval for β is the set of β_0 values for the test of H_0: $\beta = \beta_0$ such that the P-value is larger than 0.05. For example, the 95% *Wald confidence interval* is the set of β_0 values for which $z = (\hat{\beta} - \beta_0)/SE$ has $|z| < 1.96$. It is $\hat{\beta} \pm 1.96(SE)$.

1.4.2 Example: Wald, Score, and Likelihood-Ratio Binomial Tests

We illustrate the Wald, score, and likelihood-ratio tests by testing H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$ for the toy example mentioned on page 16 of a clinical trial to evaluate a new treatment that has 9 successes in $n = 10$ trials. The sample proportion is $\hat{\pi} = 0.90$.

[8] Proposed by the statistician Calyampudi Radhakrishna Rao in 1948.

[9] Proposed by the statistician Sam Wilks in 1938; in this text, we use the *natural log*, which has $e = 2.718...$ as the base. It is often denoted on calculators by LN.

For the Wald test, the estimated standard error is $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{0.90(0.10)/10} = 0.095$. The z test statistic is

$$z = (\hat{\pi} - \pi_0)/SE = (0.90 - 0.50)/0.095 = 4.22.$$

The corresponding chi-squared statistic is $(4.22)^2 = 17.78$ ($df = 1$). The P-value < 0.001.

For the score test, the null standard error is $SE_0 = \sqrt{\pi_0(1-\pi_0)/n} = \sqrt{0.50(0.50)/10} = 0.158$. The z test statistic is

$$z = (\hat{\pi} - \pi_0)/SE_0 = (0.90 - 0.50)/0.158 = 2.53.$$

The corresponding chi-squared statistic is $(2.53)^2 = 6.40$ ($df = 1$). The P-value $= 0.011$.

The likelihood function is the binomial probability of the observed result of 9 successes in 10 trials, viewed as a function of the parameter,

$$\ell(\pi) = \frac{10!}{9!1!}\pi^9(1-\pi)^1 = 10\pi^9(1-\pi).$$

The likelihood-ratio test compares this when H_0: $\pi = 0.50$ is true, for which $\ell_0 = 10(0.50)^9(0.50) = 0.00977$, to the value at the ML estimate of $\hat{\pi} = 0.90$, for which $\ell_1 = 10(0.90)^9(0.10) = 0.3874$. The likelihood-ratio test statistic equals

$$2\log(\ell_1/\ell_0) = 2[\log(0.3874/0.00977)] = 7.36.$$

From the chi-squared distribution with $df = 1$, this statistic has P-value $= 0.007$.

A marked divergence in the values of the three statistics, such as often happens when n is small and the ML estimate is near the boundary of the parameter space, indicates that the sampling distribution of the ML estimator may be far from normality and an estimate of the standard error may be poor. In that case, special small-sample methods are more reliable.

1.4.3 Small-Sample Binomial Inference and the Mid P-Value *

For statistical inference about a binomial parameter, the large-sample likelihood-ratio and two-sided score tests and the confidence intervals based on those tests perform reasonably well when $n\pi \geq 5$ and $n(1-\pi) \geq 5$. Otherwise, it is better to use the binomial distribution directly. With modern software, we can use this direct approach with any n.

To illustrate using the binomial directly, consider testing H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$ for the toy example of a clinical trial, with $y = 9$ successes in $n = 10$ trials. The exact P-value, based on the right tail of the null binomial distribution with $\pi = 0.50$, is the binomial probability

$$P(Y \geq 9) = P(9) + P(10) = \frac{10!}{9!1!}(0.50)^9(0.50)^1 + \frac{10!}{10!0!}(0.50)^{10}(0.50)^0 = 0.011.$$

For the two-sided alternative H_a: $\pi \neq 0.50$, the P-value is

$$P(Y \geq 9 \text{ or } Y \leq 1) = 2[P(Y \geq 9)] = 0.021.$$

With discrete probability distributions, small-sample inference using the ordinary P-value is *conservative*. This means that when H_0 is true, the P-value is ≤ 0.05 (thus leading to rejection of H_0 at the 0.05 significance level) not *exactly* 5% of the time, but *no more* than 5% of the time. Then, the actual P(Type I error) is not exactly 0.05, but may be much less than 0.05. For example, for testing H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$ with $y = 9$ successes in $n = 10$ trials, from the binomial probabilities with $\pi = 0.50$ in Table 1.1 in Section 1.2.1, the right-tail P-value is ≤ 0.05 only when $y = 9$ or 10. This happens with probability $0.010 + 0.001 = 0.011$. Thus, the probability of rejecting H_0 (i.e., getting a P-value ≤ 0.05) is only 0.011. That is, the actual P(Type I error) $= 0.011$, much smaller than the intended significance level of 0.05.

This illustrates an awkward aspect of small-sample significance testing when the test statistic has a discrete distribution. Imagine how a P-value, regarded as a random variable, may vary from study to study. For test statistics having a *continuous* distribution, the P-value has a *uniform* null distribution over the interval [0, 1]. That is, when H_0 is true, the P-value is equally likely to fall anywhere between 0 and 1. Then, the probability that the P-value falls below 0.05 equals exactly 0.05. The expected value of the P-value, that is, its long-run average value, is exactly 0.50. By contrast, for a test statistic having a *discrete* distribution, the null distribution of the P-value is discrete and has an expected value greater than 0.50 (e.g., it can equal 1.00 but never exactly 0.00). In this average sense, ordinary P-values for discrete distributions tend to be too large.

To address the conservatism difficulty, with discrete data we recommend using a different type of P-value. Called the *mid P-value*, it adds only *half* the probability of the observed result to the probability of the more extreme results. To illustrate, with $y = 9$ successes in $n = 10$ trials, the ordinary P-value for H_a: $\pi > 0.50$ is $P(9) + P(10) = 0.010 + 0.001 = 0.011$. The mid P-value is $[P(9)/2] + P(10) = (0.010/2) + 0.001 = 0.006$. The two-sided mid P-value for H_a: $\pi \neq 0.50$ is 0.012. The mid P-value has a null expected value of 0.50, the same as the regular P-value for test statistics that have a continuous distribution. Also, the two separate one-sided mid P-values sum to 1.0. By contrast, the observed result has probability counted in each tail for the ordinary one-sided P-values, so the two one-sided P values have a sum exceeding 1.

Inference based on the mid P-value compromises between the conservativeness of small-sample methods and the potential inadequacy of large-sample methods. It is also possible to construct a confidence interval for π from the set of π_0 values not rejected in the corresponding binomial test using the mid P-value. We shall do this with software in Section 1.6. In that section, we will see that it is straightforward to use software to obtain all the results for the examples in this chapter.

1.5 BAYESIAN INFERENCE FOR PROPORTIONS *

This book mainly uses the traditional, so-called *frequentist*, approach to statistical inference. This approach treats parameter values as fixed and data as realizations of random variables that have some assumed probability distribution. That is, probability statements refer to possible values for the data, given the parameter values. Recent years have seen increasing popularity of the *Bayesian* approach, which also treats parameters as random variables and therefore has probability distributions for them as well as for the data. This yields inferences

in the form of probability statements about possible values for the parameters, given the observed data.

1.5.1 The Bayesian Approach to Statistical Inference

The Bayesian approach assumes a *prior distribution* for the parameters. This probability distribution may reflect subjective prior beliefs, or it may reflect information about the parameter values from other studies, or it may be relatively non-informative so that inferential results are more objective, based almost entirely on the data. The prior distribution combines with the information that the data provide through the likelihood function to generate a *posterior distribution* for the parameters. The posterior distribution reflects the information about the parameters based both on the prior distribution and the data observed in the study.

For a parameter β and data denoted by y, let $f(\beta)$ denote the probability function[10] for the prior distribution of β. For example, when β is the binomial parameter π, this is a probability distribution over the interval [0, 1] of possible values for the probability π. Also, let $p(y \mid \beta)$ denote the probability function for the data, given the parameter value. (The vertical slash $\mid$ symbolizes "given" or "conditional on.") An example is the binomial formula (1.1), treating it as a function of y for fixed π. Finally, let $g(\beta \mid y)$ denote the probability function for the posterior distribution of β after we observe the data. In these symbols, from *Bayes' Theorem*,

$$g(\beta \mid y) \text{ is proportional to } p(y \mid \beta) f(\beta).$$

Now, after we observe the data, $p(y \mid \beta)$ is the likelihood function $\ell(\beta)$ when we view it as a function of the parameter. Therefore, the posterior distribution of the parameter is determined by the product of the likelihood function with the probability function for the prior distribution. When the prior distribution is relatively flat, as data analysts often choose in practice, the posterior distribution for the parameter has a similar shape to the likelihood function.

Except in a few simple cases, such as presented next for the binomial parameter, the posterior distribution cannot be easily calculated and software uses simulation methods to approximate it. The primary method for doing this is called *Markov chain Monte Carlo* (MCMC). It is beyond our scope to discuss the technical details of how an MCMC algorithm works. In a nutshell, software generates a very long sequence of values taken from an approximation for the posterior distribution. The data analyst takes the sequence to be long enough so that the Monte Carlo error is small in approximating the posterior distribution and summary measures of it, such as the mean.

For a particular parameter, Bayesian inference methods using the posterior distribution parallel those for frequentist inference. For example, analogous to the frequentist 95% confidence interval, we can construct an interval that contains 95% of the posterior distribution. Such an interval is referred to as a *posterior interval* or *credible interval*. A simple posterior interval uses percentiles of the posterior distribution, with equal probabilities in the two tails. For example, the 95% equal-tail posterior interval for a parameter is the region between the 2.5 and 97.5 percentiles of the posterior distribution. The mean of the posterior

[10] For a continuous distribution such as the normal, this is called the *probability density function*.

distribution is a Bayesian point estimator of the parameter. In lieu of P-values, posterior tail probabilities are useful, such as the posterior probability that an effect parameter in a model is positive.

1.5.2 Bayesian Binomial Inference: Beta Prior Distributions

Bayesian inference for a binomial parameter π can use a *beta distribution* as the prior distribution. The beta probability density function for π is proportional to

$$f(\pi) \propto \pi^{\alpha-1}(1-\pi)^{\beta-1}, \quad 0 \le \pi \le 1.$$

The distribution depends on two indices $\alpha > 0$ and $\beta > 0$, which are often referred to as *hyperparameters* to distinguish them from the parameter π that is the object of the inference. The mean of the beta distribution is

$$E(\pi) = \alpha/(\alpha+\beta).$$

The family of beta probability density functions has a wide variety of shapes.[11] When $\alpha = \beta$, it is symmetric around 0.50. The *uniform distribution*, $f(\pi) = 1$ over [0, 1], spreads the mass uniformly over the interval. It is the special case of a beta distribution with $\alpha = \beta = 1$. The beta density has a bimodal U-shape when $\alpha = \beta < 1$ and a bell shape when $\alpha = \beta > 1$. The variability decreases as $\alpha = \beta$ increases.

Lack of prior knowledge about π might suggest using a uniform prior distribution. The posterior distribution then has the same shape as the binomial likelihood function. Alternatively, a popular prior distribution with Bayesians is the so-called *Jeffreys prior*, for which prior distributions for different scales of measurement for the parameter (e.g., for π or for $\phi = \log[\pi/(1-\pi)]$) are equivalent. For a binomial parameter, the Jeffreys prior is the beta distribution with $\alpha = \beta = 0.5$, which has a symmetric U-shape. Although it is not flat, this prior distribution is relatively noninformative, in the sense that it has greater variability than the uniform distribution and yields inferential results similar to those of the best frequentist methods. For example, its posterior intervals have actual coverage probability close to the nominal level.[12] Unless you have reason to use something else, we recommend using it or the uniform prior distribution.

The beta distribution is the *conjugate prior distribution* for inference about a binomial parameter. This means that it is the family of probability distributions such that, when combined with the likelihood function, the posterior distribution falls in the same family. When we combine a beta(α, β) prior distribution with a binomial likelihood function, the posterior distribution is a beta distribution indexed by $\alpha^* = y + \alpha$ and $\beta^* = n - y + \beta$. The Bayesian point estimate of π is the mean of this posterior distribution,

$$\frac{\alpha^*}{\alpha^*+\beta^*} = \frac{y+\alpha}{n+\alpha+\beta} = \left(\frac{n}{n+\alpha+\beta}\right)\frac{y}{n} + \left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)\frac{\alpha}{\alpha+\beta}.$$

This estimate is a weighted average of the sample proportion $\hat{\pi} = y/n$ and the mean of the prior distribution, $\alpha/(\alpha+\beta)$. The weight $n/(n+\alpha+\beta)$ given to the sample proportion

[11] For graphs of such shapes, see `https://en.wikipedia.org/wiki/Beta_distribution`.
[12] For example, Chapter 6 of *Statistical Intervals* by W. Meeker, G. Hahn, and L. Escobar (Wiley, 2017).

increases toward 1 as n increases. With $\alpha = \beta$, the estimate shrinks the sample proportion toward 0.50. To construct the equal-tail 95% posterior interval, software uses the 2.5 and 97.5 percentiles of this posterior beta distribution.

1.5.3 Example: Opinions about Legalized Abortion, Revisited

In Section 1.3.3 we estimated the population proportion of Americans who support unrestricted legalized abortion. For a sample of $n = 1810$ people, $y = 837$ were in support and $n - y = 973$ were not. The ML estimate of π is $\hat{\pi} = 0.462$ and the 95% score confidence interval is (0.440, 0.485). How does this compare to Bayesian point and interval estimates?

For the Jeffreys beta(0.5, 0.5) prior distribution with $y = 837$ and $n - y = 973$, the posterior distribution is beta(α^*, β^*) with $\alpha^* = y + \alpha = 837.5$ and $\beta^* = n - y + \beta = 973.5$. The posterior mean estimate of π is $\alpha^*/(\alpha^* + \beta^*) = 837.5/(837.5 + 973.5) = 0.462$. Software (e.g., as shown in Section 1.6.2) reports the posterior 95% equal-tail interval of (0.440, 0.485), the endpoints being the 2.5 and 97.5 percentiles of the beta posterior density.

The Bayesian point estimate and posterior interval are the same, to three decimal places, as the ML estimate and frequentist 95% score interval. Frequentist and Bayesian inferences tend to be very similar when n is large and the prior distribution is highly disperse. However, the interpretations are quite different. With the frequentist approach, the actual parameter value π either *is* or *is not* in the confidence interval of (0.440, 0.485). Our 95% confidence means that if we used this method over and over with separate, independent samples, in the long run 95% of the confidence intervals would contain π. That is, the probability applies to possible data in future samples, not to the parameter. By contrast, with the Bayesian approach, after observing the data, we can say that the probability is 0.95 that π falls between 0.440 and 0.485.

The ordinary frequentist P-value for the score test (Section 1.3.2) of H_0: $\pi = 0.50$ against H_a: $\pi < 0.50$ is 0.000695. For such a one-sided test, the implicit null hypothesis is H_0: $\pi \geq 0.50$, and we use the boundary value to form the test statistic. A corresponding Bayesian posterior probability of interest is $P(\pi \geq 0.50)$, which equals 0.000692. The frequentist interpretation is that if H_0 were true (i.e., if $\pi = 0.50$), the probability of getting a test statistic like the observed one or even more extreme in the direction of H_a is 0.000695. This is a probability about potential data, given a parameter value. By contrast, the Bayesian interpretation of the posterior probability is that, after observing the data, the probability that $\pi \geq 0.50$ is 0.000692. With highly disperse prior distributions, such a one-tail probability is approximately equal to the frequentist one-sided P-value.

1.5.4 Other Prior Distributions

Bayesian methods for binomial parameters can use prior distributions other than the beta distribution. One possibility, hierarchical in nature, also assumes prior distributions for the beta hyperparameters instead of assigning fixed values.

For a prior distribution for $c > 2$ multinomial parameters, the beta distribution generalizes to the *Dirichlet distribution*. It is defined over the simplex of nonnegative values $(\pi_1, \ldots, \pi_c)$ that sum to 1. The posterior distribution is then also Dirichlet.

Models presented in this book have effect parameters that can take any real-number value. Bayesian methods for such parameters typically use normal prior distributions.

1.6 USING R SOFTWARE FOR STATISTICAL INFERENCE ABOUT PROPORTIONS *

Statistical software packages can implement categorical data analyses. The software package R is increasingly popular, partly because it is available to download for free at www.r-project.org and partly because users can contribute their own functions to make new methods available. Throughout the text, we show the R code and output for the examples. The Appendix shows examples for SAS, Stata, and SPSS software.

You can get help about R at many sites on the Internet, such as https://stats.idre.ucla.edu/r. Many users prefer to use RStudio, an integrated development environment (IDE) for R. It includes a code editor and tools for debugging and plotting. See www.rstudio.com and community.rstudio.com.

For a particular function or command, using R you can get help by placing a ? before its name, for example, entering

```
> ?prop.test
```

to get information about the function prop.test that can perform inference about proportions.

1.6.1 Reading Data Files and Installing Packages

In this book, we show R statements that request statistical analyses from the command line. A basic command loads a data file from the text website,[13] www.stat.ufl.edu/~aa/cat. For example, the data file called Clinical.dat at the text website has the observations for the clinical trials toy example in Section 1.3.5, with 9 successes in 10 observations. Here is how to create an R data file with the name Clinical from that data file at the text website and then view the file:

```
------------------------------------------------------------------------------
> Clinical <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Clinical.dat",
+                        header=TRUE)
> Clinical
   subject response
1        1        1
2        2        1
3        3        1
4        4        1
5        5        1
6        6        1
7        7        1
8        8        1
9        9        1
10      10        0
------------------------------------------------------------------------------
```

The *header=TRUE* part of the read.table command tells R that the first row of the data file contains the variable names. The standard form for a data file has a separate row for each subject (e.g., person) in the sample and a separate column for each variable. Here, the

[13] You can also copy the data files from https://github.com/alanagresti/categorical-data.

column labelled *response* shows the 0 and 1 values for failure and success indications of a binary outcome.

Many users of R have created packages that can perform analyses not available in basic R. For example, to install the binom package that can conduct some statistical inference for binomial parameters, use the command

```
> install.packages("binom")
```

Once it is installed, load the package to access its functions:

```
> library(binom)
```

1.6.2 Using R for Statistical Inference about Proportions

For the opinions about legalized abortion example (Section 1.3.3) with a binomial count of 837 supporting unrestricted legalization out of $n = 1810$, here is how you can conduct two-sided and one-sided significance tests and a confidence interval for the population proportion:

```
------------------------------------------------------------------------------
> prop.test(837, 1810, p=0.50, alternative="two.sided", correct=FALSE)
data:  837 out of 1810, null probability 0.5
X-squared = 10.219, df = 1, p-value = 0.00139 # chi-squared, 2-sided altern.
alternative hypothesis: true p is not equal to 0.5 # "true p" is binom. para.
95 percent confidence interval:# score CI
 0.4395653 0.4854557

> prop.test(837, 1810, p=0.50, alternative="less", correct=FALSE)
data:  837 out of 1810, null probability 0.5
X-squared = 10.219, df = 1, p-value = 0.000695 # chi-squared, 1-sided altern.
alternative hypothesis: true p is less than 0.5
------------------------------------------------------------------------------
```

The *X-squared* value in the R output is the *score* test statistic, having a chi-squared null distribution with $df = 1$. The z test statistic (1.2) that we presented is the positive or negative square root of this value. The *correct=FALSE* option stops R from using a *continuity correction*. We do not recommend using continuity corrections, because inferences then tend to be too conservative. The confidence interval displayed is the *score* confidence interval introduced in Section 1.3.5.

We can also request analyses directly from a data file. Suppose we load a data file in which one column contains 0 and 1 values for failure and success indications of a binary outcome. We can define a binomial variable that sums these indicators to obtain the number of successes and then conduct inferences using it. For example, for the Clinical data file for the 9 successes in 10 observations for the clinical trials example in Section 1.3.5, here is how we could find the score confidence interval quoted in that example:

```
------------------------------------------------------------------------------
> Clinical
   subject response
1        1        1
2        2        1
```

```
...
10      10        0
> attach(Clinical)
> y <- sum(response) # sums 0 and 1 values to get number of successes
> prop.test(y, n=10, conf.level=0.95, correct=FALSE)
95 percent confidence interval:
 0.59585  0.98212 # score CI for probability of success
------------------------------------------------------------------------
```

Rather than attach a data file, which can cause confusion if a variable has already been defined in the R session that has the same name as a variable in that data file, you can refer to the data file name in the command itself or you can instead imbed the data file name and desired command in a `with` function:

```
------------------------------------------------------------------------
> prop.test(sum(Clinical$response), 10, correct=FALSE)$conf.int
> with(Clinical, prop.test(sum(response), 10, correct=FALSE)$conf.int)
------------------------------------------------------------------------
```

The Wald confidence interval ("asymptotic"), score interval ("wilson," named after the statistician who first proposed this interval), and Agresti–Coull interval (near the end of Section 1.3.5) are available with the `binom` package. Here we show them for 9 successes in 10 trials:

```
------------------------------------------------------------------------
> library(binom)
> binom.confint(9, 10, conf.level=0.95, method="asymptotic")
      method x  n  mean    lower    upper
  asymptotic 9 10   0.9  0.71406  1.08594 # Wald confidence interval
> binom.confint(9, 10, conf.level=0.95, method="wilson")
  method x  n  mean    lower    upper
  wilson 9 10   0.9  0.59585  0.98212 # score confidence interval
> binom.confint(9, 10, conf.level=0.95, method="agresti-coull")
         method x  n  mean    lower    upper
  agresti-coull 9 10   0.9  0.57403  1.00394
------------------------------------------------------------------------
```

For any upper bound reported above 1.0, you should truncate it at 1.000, since π must fall between 0 and 1. Likelihood-ratio tests and corresponding confidence intervals are easy to obtain in R in the context of modeling, as we will see in future chapters.

The `binom.test` function uses the binomial distribution to obtain exact P-values. For example, with $y = 9$ and $n = 10$, we find the P-value for H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$ and against H_a: $\pi > 0.50$:

```
------------------------------------------------------------------------
> binom.test(9, 10, 0.50, alternative = "two.sided")
        Exact binomial test
number of successes = 9, number of trials = 10, p-value = 0.02148
alternative hypothesis: true probability of success is not equal to 0.5
```

```
> binom.test(9, 10, 0.50, alternative = "greater")
        Exact binomial test
number of successes = 9, number of trials = 10, p-value = 0.01074
alternative hypothesis: true probability of success is greater than 0.5
------------------------------------------------------------------------------
```

Binomial tests using the mid P-value are available with the `exactci` package. Binomial confidence intervals using the mid P-value are available with that package and with the `PropCIs` package. Again, here are results for $y = 9$ in $n = 10$ trials:

```
------------------------------------------------------------------------------
> library(exactci)
> binom.exact(9, 10, 0.50, alternative="greater", midp=TRUE) # mid P-value
number of successes = 9, number of trials = 10, p-value = 0.005859
> library(PropCIs)
> midPci(9, 10, 0.95) # confidence interval based on test with mid P-value
 0.5966 0.9946
------------------------------------------------------------------------------
```

A Bayesian posterior interval based on a beta(α, β) prior distribution and y successes and $n - y$ failures finds percentiles of the beta distribution with parameters $\alpha^* = y + \alpha$ and $\beta^* = n - y + \beta$ using the `qbeta` quantile function. We find a 95% posterior interval here using $\alpha = \beta = 0.5$ with $y = 837$ and $n - y = 973$, as in the opinion about legalized abortion example (Section 1.5.3). We can use the `pbeta` cumulative probability function to find a tail probability, such as the posterior $P(\pi \geq 0.50)$:

```
------------------------------------------------------------------------------
> qbeta(c(0.025, 0.975), 837.5, 973.5)
[1] 0.43954  0.48545 # bounds of posterior interval, for Jeffreys prior

> pbeta(0.50, 837.5, 973.5) # posterior beta cumulative prob. at 0.50
[1] 0.99931
> 1 - pbeta(0.50, 837.5, 973.5) # right-tail probability above 0.50
[1] 0.00069 # can also get as pbeta(0.50, 837.5, 973.5, lower.tail=FALSE)
------------------------------------------------------------------------------
```

1.6.3 Summary: Choosing an Inference Method

In summary, several methods are available for conducting inference about a binomial parameter. It can be confusing for a methodologist to decide which to use. With modern computing power, it is no longer necessary in this simple setting to rely on methods based on large-sample normal or chi-squared approximations. For a frequentist approach to significance testing or confidence intervals, we recommend exact binomial inference using the mid P-value. For the Bayesian approach, we recommend inference using the beta posterior distribution induced by a beta(0.5, 0.5) prior distribution.

EXERCISES

1.1 In the following examples, identify the natural response variable and the explanatory variables.

a. Attitude toward gun control (favor, oppose), gender (female, male), mother's education (high school, college).

b. Heart disease (yes, no), blood pressure, cholesterol level.

c. Race (white, nonwhite), religion (Catholic, Jewish, Muslim, Protestant, none), vote for president (Democrat, Republican, Green), annual income.

1.2 Which scale of measurement is most appropriate for the following variables — nominal or ordinal?

a. UK political party preference (Labour, Liberal Democrat, Conservative, other)

b. Highest educational degree obtained (none, high school, bachelor's, master's, doctorate).

c. Patient condition (good, fair, serious, critical).

d. Hospital location (London, Boston, Madison, Rochester, Toronto).

e. Favorite beverage (beer, juice, milk, soft drink, wine, other).

f. Rating of a movie with 1 to 5 stars, representing (hated it, didn't like it, liked it, really liked it, loved it)

1.3 Each of 100 multiple-choice questions on an exam has four possible answers but one correct response. For each question, a student randomly selects one response as the answer.

a. Specify the probability distribution of the student's number of correct answers on the exam.

b. Based on the mean and standard deviation of that distribution, would it be surprising if the student made at least 50 correct responses? Explain your reasoning.

1.4 In a particular city, the population proportion π supports an increase in the minimum wage. For a random sample of size 2, let $Y =$ number who support an increase.

a. Assuming $\pi = 0.50$, specify the probabilities for the possible values y for Y and find the distribution's mean and standard deviation.

b. Suppose you observe $y = 1$ and do not know π. Find and sketch the likelihood function. Using the plotted likelihood function, explain why the ML estimate $\hat{\pi} = 0.50$.

1.5 Refer to the previous exercise. Suppose $y = 0$ for $n = 2$. Find the ML estimate of π. Does this estimate seem believable? Why? Find the Bayesian estimator based on the prior belief that π is equally likely to be anywhere between 0 and 1.

1.6 Genotypes AA, Aa, and aa occur with probabilities (π_1, π_2, π_3). For $n = 3$ independent observations, the observed frequencies are (y_1, y_2, y_3).

a. Explain how you can determine y_3 from knowing y_1 and y_2. Thus, the multinomial distribution of (y_1, y_2, y_3) is actually two-dimensional.

b. Show the ten possible observations (y_1, y_2, y_3) with $n = 3$.

c. Suppose $(\pi_1, \pi_2, \pi_3) = (0.25, 0.50, 0.25)$. What probability distribution does y_1 alone have?

1.7 In his autobiography *A Sort of Life*, British author Graham Greene described a period of severe mental depression during which he played Russian Roulette — putting a bullet in one of the six chambers of a pistol, spinning the chambers to select one at random, and then firing the pistol once at one's head.

a. Greene played this game six times, and was lucky that none of them resulted in a bullet firing. Find the probability of this outcome.

b. Suppose he had kept playing this game until the bullet fires. Let Y denote the number of the game on which the bullet fires. Explain why the probability of the outcome y equals $(5/6)^{y-1}(1/6)$, for $y = 1, 2, 3, \ldots$. (This is called the *geometric distribution*.)

1.8 When the 2010 General Social Survey asked subjects in the US whether they would be willing to accept cuts in their standard of living to protect the environment, 486 of 1374 subjects said *yes*.

a. Estimate the population proportion who would say *yes*. Construct and interpret a 99% confidence interval for this proportion.

b. Conduct a significance test to determine whether a majority or minority of the population would say *yes*. Report and interpret the P-value.

1.9 A study of 100 women suffering from excessive menstrual bleeding considers whether a new analgesic provides greater relief than the standard analgesic. Of the women, 40 reported greater relief with the standard analgesic and 60 reported greater relief with the new one.

a. Test the hypothesis that the probability of greater relief with the standard analgesic is the same as the probability of greater relief with the new analgesic. Report and interpret the P-value for the two-sided alternative. (*Hint:* Express the hypotheses in terms of a single parameter. A test to compare matched-pairs responses in terms of which is better is called a *sign test*.)

b. Construct and interpret a 95% confidence interval for the probability of greater relief with the new analgesic.

1.10 Refer to the previous exercise. The researchers wanted a sufficiently large sample to be able to estimate the probability of preferring the new analgesic to within 0.08, with confidence 0.95. If the true probability is 0.75, how large a sample is needed to achieve this accuracy? (*Hint:* For how large an n does a 95% confidence interval have margin of error equal to about 0.08?)

1.11 When a recent General Social Survey asked 1158 American adults, "Do you believe in heaven?", the proportion who answered yes was 0.86. Treating this as a random sample, conduct statistical inference about the population proportion of American adults believing in heaven. Summarize your analysis and interpret the results in a short report.

1.12 To collect data in an introductory statistics course, I gave the students a questionnaire. One question asked whether the student was a vegetarian. Of 25 students, 0 answered

yes. They were not a random sample, but use these data to illustrate inference for a proportion. Let π denote the population proportion who would say *yes*. Consider H_0: $\pi = 0.50$ and H_a: $\pi \neq 0.50$.

a. What happens when you conduct the *Wald test*, which uses the *estimated* standard error in the z test statistic?

b. Find the 95% *Wald confidence interval* (1.3) for π. Is it believable?

c. Conduct the *score test*, which uses the *null* standard error in the z test statistic. Report and interpret the P-value.

d. Verify that the 95% score confidence interval equals (0.0, 0.133). (This is similar to the interval (0.0, 0.137) obtained with a small-sample method of Section 1.4.3, inverting the binomial test with the mid P-value.)

1.13 Refer to the previous exercise, with $y = 0$ in $n = 25$ trials for testing H_0: $\pi = 0.50$.

a. Show that ℓ_0, the maximized likelihood under H_0, equals $(1 - \pi_0)^{25} = (0.50)^{25}$. Show that ℓ_1, the maximized likelihood over all possible π values, equals 1.0. (*Hint:* This is the value at the ML estimate value of 0.0.)

b. Show that the likelihood-ratio test statistic, $2\log(\ell_1/\ell_0)$, equals 34.7. Report the P-value.

c. The 95% likelihood-ratio-test-based confidence interval for π is (0.000, 0.074). Verify that 0.074 is the correct upper bound by showing that the likelihood-ratio test of H_0: $\pi = 0.074$ against H_a: $\pi \neq 0.074$ has a chi-squared test statistic equal to 3.84 and P-value = 0.05.

1.14 Section 1.4.3 found binomial P-values for a clinical trial with $y = 9$ successes in 10 trials. Suppose instead $y = 8$. Using software or the binomial distribution shown in Table 1.1:

a. Find the P-value for (i) H_a: $\pi > 0.50$, (ii) H_a: $\pi < 0.50$.

b. Find the mid P-value for (i) H_a: $\pi > 0.50$, (ii) H_a: $\pi < 0.50$.

c. Why is the sum of the one-sided P-values greater than 1.0 for the ordinary P-value but equal to 1.0 for the mid P-value?

d. Using software, find the 95% confidence interval based on the binomial test with the mid P-value.

1.15 If Y is a random variable and c is a positive constant, then the standard deviation of the probability distribution of cY equals $c\sigma(Y)$. Suppose Y is a binomial variate and let $\hat{\pi} = Y/n$.

a. Based on the binomial standard deviation for Y, show that $\sigma(\hat{\pi}) = \sqrt{\pi(1-\pi)/n}$.

b. Explain why it is easier to estimate π precisely when it is near 0 or 1 than when it is near 0.50.

1.16 Using calculus, it is easier to derive the maximum of the log of the likelihood function, $L = \log \ell$, than the likelihood function ℓ itself. Both functions have a maximum at the same value, so it is sufficient to do either.

a. Calculate the log-likelihood function $L(\pi)$ for the binomial distribution (1.1).

b. One can usually determine the point at which the maximum of a log-likelihood L occurs by solving the *likelihood equation*. This is the equation resulting from

differentiating L with respect to the parameter and setting the derivative equal to zero. Find the likelihood equation for the binomial distribution and solve it to show that the ML estimate is $\hat{\pi} = y/n$.

1.17 Refer to Exercise 1.12 on estimating the population proportion π of vegetarians. For the beta(0.5, 0.5) prior, find the Bayes estimator of π, show that the posterior 95% interval is (0.00002, 0.0947), and show that the posterior $P(\pi < 0.50) = 1.000$.

1.18 For the previous exercise, explain how the Bayes estimator shrinks the sample proportion toward the prior mean.

1.19 For Exercises 1.12 and 1.17, explain the difference between the frequentist interpretation of the score confidence interval (0.0, 0.133) and the Bayesian interpretation of the posterior interval (0.00002, 0.0947).

CHAPTER 2

ANALYZING CONTINGENCY TABLES

Table 2.1 cross-classifies a random sample of Americans according to their gender and their belief in an afterlife. For the 1587 females in the sample, for example, 1230 said they believed in an afterlife and 357 said they did not or were undecided. Does an association exist between gender and belief in an afterlife? Is one gender more likely than the other to believe in an afterlife, or is belief in an afterlife plausibly independent of gender?

Table 2.1 Cross-classification of belief in afterlife by gender.

	Belief in Afterlife		
Gender	Yes	No or Undecided	Total
Females	1230	357	1587
Males	859	413	1272
Total	2089	770	2859

Source: Data from 2016 General Social Survey.

Analyzing associations is at the heart of multivariate statistical analysis. This chapter introduces parameters that describe the association between categorical variables and presents inferential methods for those parameters.

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

2.1 PROBABILITY STRUCTURE FOR CONTINGENCY TABLES

For a rectangular table that cross-classifies categorical variables X and Y, we let r denote the number of categories of X (the rows) and c the number of categories of Y (the columns). This table has *cells* that display the rc possible combinations of outcomes.

A table of this form that displays counts of outcomes in the cells is called a *contingency table*. A table that cross-classifies two variables is called a *two-way contingency table*; one that cross-classifies three variables is called a *three-way contingency table*, and so forth. A two-way table with r rows and c columns is called an $r \times c$ (read as r–by–c) table. Table 2.1 is a 2×2 table.

2.1.1 Joint, Marginal, and Conditional Probabilities

Probabilities for contingency tables can be of three types – *joint*, *marginal*, or *conditional*. Suppose first that a randomly chosen subject from the population of interest is classified on X and Y. Let $\pi_{ij} = P(X = i, Y = j)$ denote the probability that (X, Y) falls in the cell in row i and column j. The probabilities $\{\pi_{ij}\}$ form the *joint distribution* of X and Y. They satisfy $\sum \pi_{ij} = 1$, with the sum taken over all rc cells. We use corresponding estimator notation for samples, with $\{\hat{\pi}_{ij}\}$ denoting cell proportions in a sample joint distribution. For cell counts $\{n_{ij}\}$, with $n = \sum n_{ij}$ being the total sample size,

$$\hat{\pi}_{ij} = n_{ij}/n.$$

In Table 2.1, for example, of the $n = 2859$ respondents who are cross-classified by their gender and by their belief in an afterlife, $n_{11} = 1230$, and the related sample joint proportion is $\hat{\pi}_{11} = 1230/2859 = 0.43$. Likewise, $\hat{\pi}_{12} = 0.12$, $\hat{\pi}_{21} = 0.30$, and $\hat{\pi}_{22} = 0.14$.

The *marginal distributions* are the row and column totals of the joint probabilities. We denote these by $\{\pi_{i+}\}$ for the row variable and $\{\pi_{+j}\}$ for the column variable, where the subscript "+" denotes the sum over the index it replaces. For 2×2 tables,

$$\pi_{1+} = \pi_{11} + \pi_{12} \quad \text{and} \quad \pi_{+1} = \pi_{11} + \pi_{21}$$

are the probabilities for row 1 and for column 1. Each marginal distribution refers to a single variable. Table 2.1 has sample marginal distribution for belief in an afterlife of (2089/2859, $770/2859$) = (0.73, 0.27) for the categories (yes, no, or undecided).

In two-way tables, usually one variable (say, the column variable, Y) is a response variable and the other (the row variable, X) is an explanatory variable. Then, it is informative to construct a separate probability distribution for Y at each value of X. Such a distribution consists of *conditional probabilities* for Y, given X. It is called a *conditional distribution*. In Table 2.1, belief in the afterlife is a response variable and gender is an explanatory variable. Therefore, we form the conditional distributions of belief in the afterlife, given gender. For females, the proportion of *yes* responses was $1230/1587 = 0.78$ and the proportion of *no or undecided* responses was $357/1587 = 0.22$. The proportions (0.78, 0.22) form the sample conditional distribution of belief in the afterlife. For males, the sample conditional distribution is (0.68, 0.32).

2.1.2 Example: Sensitivity and Specificity

Diagnostic testing is used in medicine to detect many diseases. For example, the mammogram can detect breast cancer in women, and the Prostate-Specific Antigen (PSA) test can

detect prostate cancer in men. The outcome of a diagnostic test is said to be *positive* if it predicts that the disease is present and *negative* if it predicts that the disease is absent.

Let X denote the true state of a person (1 = diseased, 2 = not diseased), and let Y = outcome of diagnostic test (1 = positive, 2 = negative). The accuracy of diagnostic tests is assessed with two conditional probabilities. Given that a subject has the disease ($X = 1$), the probability the diagnostic test is positive ($Y = 1$) is called the *sensitivity*. Given that the subject does *not* have the disease ($X = 2$), the probability the test is negative ($Y = 2$) is called the *specificity*. That is,

$$\text{Sensitivity} = P(Y = 1 \mid X = 1), \text{ Specificity} = P(Y = 2 \mid X = 2).$$

The higher the sensitivity and specificity, the better the diagnostic test.

In practice, if you get a positive diagnosis, the more relevant conditional probability is $P(X = 1 \mid Y = 1)$, called the *positive predictive value* (PPV). Given that the diagnostic test predicts that you have the disease, this is the probability that you truly have it. When relatively few people have the disease, this probability can be low even when the sensitivity and specificity are high. For example, typical values reported in the medical literature for mammograms are sensitivity = 0.86 and specificity = 0.88. Of women who get mammograms at any given time, it has been estimated that about 1% truly have breast cancer. For these values, given that a mammogram has a positive diagnosis, the probability[1] that the woman truly has breast cancer is only 0.07. How can the PPV be so low, given relatively high sensitivity and specificity?

Figure 2.1 explains this, using a tree diagram that shows results for a typical sample of 100 women. The first set of branches shows whether a woman has breast cancer. Here, 1 of the 100 women have it, 1% of the sample. The second set of branches shows the mammogram result, given the disease status. For a woman with breast cancer, there is a 0.86 probability of detecting it. Therefore, we would expect the 1 woman with breast cancer to have a positive diagnostic test result, as the figure shows. For a woman without breast cancer, there is a 0.88 probability of a negative result. Therefore, we would expect about

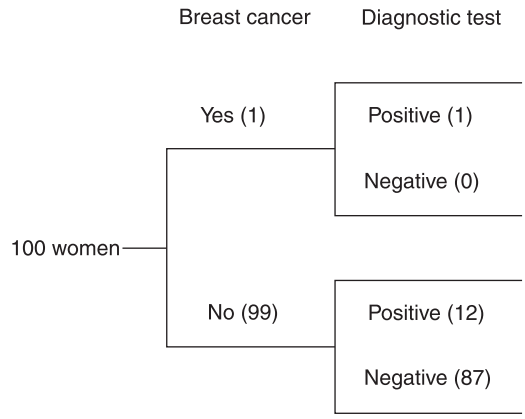


Figure 2.1 Tree diagram showing results of 100 sample mammograms, when population sensitivity = 0.86 and specificity = 0.88.

[1] This can be shown with Bayes' theorem (Exercise 2.2).

$(0.88)99 = 87$ of the 99 women without breast cancer to have a negative diagnostic test result and $(0.12)99 = 12$ to have a positive diagnostic test result. Figure 2.1 shows that of the 13 women with a positive diagnostic test result, the proportion $1/13 = 0.08$ actually have breast cancer. For a population, the low proportion of errors for the great majority of women who do not have breast cancer swamps the high proportion of correct diagnoses for the women who have it.

2.1.3 Statistical Independence of Two Categorical Variables

Two categorical variables X and Y that are *both* response variables are said to be *statistically independent* if all their joint probabilities equal the product of their marginal probabilities,

$$P(X = i,\ Y = j) = P(X = i)P(Y = j) \ \text{ for } i = 1, \ldots, r \quad \text{and} \quad j = 1, \ldots, c.$$

That is, $\pi_{ij} = \pi_{i+}\pi_{+j}$ in all cells. The marginal probabilities determine the joint probabilities.

When Y is a response variable and X is explanatory, it is more relevant to analyze the conditional distributions of Y given X. The variables are then said to be *statistically independent* if the true conditional distributions of Y are identical at each level of X. Statistical independence is then also referred to as *homogeneity* of the conditional distributions. When the variables are independent, the probability of any particular column outcome j is the same in each row. Belief in an afterlife is independent of gender, for instance, if the probability of believing in an afterlife equals 0.73 both for females and for males.

2.1.4 Binomial and Multinomial Sampling

Section 1.2 introduced the *binomial* and *multinomial* distributions. With random sampling or randomized experiments, it is common to assume that cell counts in contingency tables have one of these distributions.

When the rows of a contingency table refer to different groups, the sample sizes for those groups are often fixed by the sampling design. An example is a randomized experiment to compare a new drug to placebo in treating some illness, in which half the sample is randomly allocated to each of two treatments. When the marginal totals for X are fixed rather than random, a joint distribution for X and Y is not meaningful, but conditional distributions for Y (given X) are. When Y has two outcome categories (e.g., success or failure for response to a treatment), the binomial distribution applies for each conditional distribution. We assume a binomial distribution for the counts in each row, with the number of trials equal to the fixed row total. When Y has more than two outcome categories (e.g., complete success, partial success, or failure), the multinomial distribution applies for the counts in each row.

In most sample surveys, such as the General Social Survey, the overall sample size n is fixed. When we cross-classify a random sample on two categorical response variables X and Y, the multinomial distribution describes the joint distribution over the cells. The cells of the contingency table are the possible outcomes and the cell probabilities are the multinomial parameters. However, when Y is a response variable and X (the row variable) is an explanatory variable, it is sensible to analyze sample conditional distributions on Y,

inherently treating the row totals as fixed and analyzing the data as if the rows formed separate samples. In Table 2.1, since belief in an afterlife is the response variable, we could treat the results for females as a binomial sample with outcome categories *yes* and *no or undecided* for belief in an afterlife, and the results for males as a separate, independent binomial sample.

2.2 COMPARING PROPORTIONS IN 2 × 2 CONTINGENCY TABLES

Many studies compare two groups on a binary response variable, Y. The data can be displayed in a 2×2 contingency table, in which the rows are the groups and the columns are the response categories for Y. This section presents measures for comparing the groups.

2.2.1 Difference of Proportions

We use the generic terms *success* and *failure* for the outcome categories. Denote the probability of success by π_1 for subjects in row 1 and by π_2 for subjects in row 2. These are conditional probabilities. The probabilities of failure are $1 - \pi_1$ and $1 - \pi_2$.

The difference $\pi_1 - \pi_2$ compares the success probabilities for the two groups. It falls between -1 and $+1$, equaling zero when $\pi_1 = \pi_2$, that is, when the response variable is independent of the group classification. Let $\hat{\pi}_1$ and $\hat{\pi}_2$ denote the *sample* proportions of successes. The sample *difference of proportions* $\hat{\pi}_1 - \hat{\pi}_2$ estimates $\pi_1 - \pi_2$.

For sample sizes n_1 and n_2 for the two groups, when we treat the two samples as independent binomial samples, the estimated standard error of $\hat{\pi}_1 - \hat{\pi}_2$ is

$$SE = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}. \tag{2.1}$$

As the sample sizes increase, the standard error decreases and the estimate of $\pi_1 - \pi_2$ tends to improve. A large-sample $100(1-\alpha)\%$ Wald confidence interval for $\pi_1 - \pi_2$ is

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2}(SE).$$

For a significance test of H_0: $\pi_1 = \pi_2$, the standard z test statistic divides $(\hat{\pi}_1 - \hat{\pi}_2)$ by a pooled SE that applies under H_0. Its square is a special case of the Pearson chi-squared statistic presented in Section 2.4, so we will not present this test here.

For small samples, other confidence intervals perform better than the Wald interval, having actual coverage probability closer to the nominal confidence level. This is especially true when π_1 and π_2 are close to 0 or 1. A good general-purpose method is based on correspondence with results of a score test about $\pi_1 - \pi_2$. It is beyond our scope to explain computations here, but this interval is available in software (see Section 2.3.3). Alternatively, a simple fix of the Wald formula that improves its performance substantially is to add 1.0 to every cell of the 2×2 table before applying it[2].

[2] A. Agresti and B. Caffo, *The American Statistician* **54**: 280–288 (2000), available with the `wald2ci` function (option "AC") in the `PropCIs` package in `R`.

2.2.2 Example: Aspirin and Incidence of Heart Attacks

Table 2.2 comes from the Physicians' Health Study,[3] which was a five-year randomized study investigating whether regular intake of aspirin reduces the chance of myocardial infarction (heart attacks). Every other day, the male physicians participating in the study took either one aspirin tablet or a placebo. The study was *blind* – the physicians in the study did not know which type of pill they were taking. The later Nurses' Health Study[4] investigated the association between aspirin use and various types of cancer.

Table 2.2 Cross-classification of aspirin use and myocardial infarction (MI).

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

Source: Based on data in *N. Engl. J. Med.* **318**: 262–264 (1988).

We treat the two rows in Table 2.2 as independent binomial samples. The sample proportions of physicians that suffered myocardial infarction (MI) during the study were $\hat{\pi}_1 = 189/11,034 = 0.0171$ for those taking placebo and $\hat{\pi}_2 = 104/11,037 = 0.0094$ for those taking aspirin. The sample difference of proportions is $\hat{\pi}_1 - \hat{\pi}_2 = 0.0171 - 0.0094 = 0.0077$. From (2.1), $\hat{\pi}_1 - \hat{\pi}_2$ has an estimated standard error of

$$SE = \sqrt{\frac{(0.0171)(0.9829)}{11,034} + \frac{(0.0094)(0.9906)}{11,037}} = 0.0015.$$

A 95% confidence interval[5] for $\pi_1 - \pi_2$ is $0.0077 \pm 1.96(0.0015)$, which is 0.008 ± 0.003, or (0.005, 0.011). Since this interval contains only positive values, we conclude that $\pi_1 - \pi_2 > 0$, that is, $\pi_1 > \pi_2$. Taking aspirin corresponds to a diminished risk of heart attack.

2.2.3 Ratio of Proportions (Relative Risk)

A difference between two proportions of a certain fixed size usually is more important when both proportions are near 0 or 1 than when they are near the middle of the range. Consider a comparison of two drugs on the proportion of subjects who had adverse reactions when using them. The difference between 0.010 and 0.001 is the same as the difference between 0.410 and 0.401, namely 0.009. The first difference is more striking, since ten times as many subjects had adverse reactions with one drug as the other. In such cases, the *ratio* of proportions is a more relevant descriptive measure.

For 2×2 tables, the ratio of probabilities is often called the *relative risk*:

$$\text{Relative risk} = \frac{\pi_1}{\pi_2}.$$

It can be any nonnegative real number. The proportions 0.010 and 0.001 have a relative risk of $0.010/0.001 = 10.0$, whereas the proportions 0.410 and 0.401 have a relative risk of

[3] See `phs.bwh.harvard.edu`.

[4] See `www.nurseshealthstudy.org`.

[5] We also obtain (0.005, 0.011) with the score confidence interval, in Section 2.3.3.

$0.410/0.401 = 1.02$. A relative risk of 1.00 occurs when $\pi_1 = \pi_2$, that is, when the response variable is independent of the group. The ratio of failure probabilities, $(1 - \pi_1)/(1 - \pi_2)$, takes a different value than the ratio of the success probabilities.

For Table 2.2, the sample relative risk is $\hat{\pi}_1/\hat{\pi}_2 = 0.0171/0.0094 = 1.82$. The sample proportion of MI cases was 82% higher for the group taking a placebo. The sample difference of proportions of 0.0077 makes it seem as if the two groups differ by a trivial amount. By contrast, the relative risk shows that the difference may have important public health implications. Using the difference of proportions alone to compare two groups can be misleading when the proportions are both close to zero.

The sample relative risk has a sampling distribution that is highly skewed unless the sample sizes are quite large. Because of this, its confidence interval formula is rather complex. For Table 2.2, software (see the output below) reports a 95% confidence interval for the true relative risk of (1.43, 2.30). We infer that, after five years, the probability of MI for physicians taking placebo is between 1.43 and 2.30 times the probability of MI for physicians taking aspirin. Therefore, the risk of MI is at least 43% higher for the placebo group.

2.2.4 Using R for Comparing Proportions in 2 × 2 Tables

Here is how R can find the Wald confidence interval for $\pi_1 - \pi_2$, illustrating for the data on aspirin use and MI (Section 2.2.2):

```
------------------------------------------------------------------------------
> prop.test(c(189, 104), c(11034, 11037), conf.level=0.95, correct=FALSE)
   # uses the two success counts and n1 and n2, no continuity correction
95 percent confidence interval:
 0.00469  0.01072  # Wald CI for difference of proportions
------------------------------------------------------------------------------
```

For binomial proportions, the R package `PropCIs` has many confidence interval functions, including score confidence intervals for $\pi_1 - \pi_2$ and the relative risk:

```
------------------------------------------------------------------------------
> library(PropCIs) # uses binomial success count and n for each group
> diffscoreci(189, 11034, 104, 11037, conf.level=0.95)
95 percent confidence interval:
 0.00472  0.01079  # score CI for difference of proportions
> riskscoreci(189, 11034, 104, 11037, conf.level=0.95)
95 percent confidence interval:
 1.43390  2.30471  # score CI for relative risk
------------------------------------------------------------------------------
```

2.3 THE ODDS RATIO

The *odds ratio* is another measure of association for 2×2 contingency tables. It also occurs as a parameter in the most important model for categorical responses — logistic regression.

For a probability of success π, the *odds* of success are defined to be

$$\text{odds} = \pi/(1 - \pi).$$

For instance, when $\pi = 0.75$, the odds of success equal 0.75/0.25 = 3.0. The odds are nonnegative, with value greater than 1.0 when a success is more likely than a failure. When odds = 3.0, we expect to observe three successes for every one failure. When odds $= 1/3$, a failure is three times as likely as a success. We then expect to observe one success for every three failures.

The success probability is the function of the odds,

$$\pi = \text{odds}/(\text{odds} + 1).$$

For instance, when odds = 4, then $\pi = 4/(4 + 1) = 0.80$. The probability of success is 0.80, the probability of failure is 0.20, and the odds equal 0.80/0.20 = 4.0.

In 2×2 tables, within row 1 the *odds* of success are $\text{odds}_1 = \pi_1/(1 - \pi_1)$, and within row 2 the odds of success equal $\text{odds}_2 = \pi_2/(1 - \pi_2)$. The ratio of the odds from the two rows,

$$\theta = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)},$$

is the *odds ratio*. Whereas the relative risk is a ratio π_1/π_2 of two *probabilities*, the odds ratio θ is a ratio of two *odds*.

2.3.1 Properties of the Odds Ratio

The odds ratio can equal any nonnegative number. When X and Y are independent, $\pi_1 = \pi_2$, so $\text{odds}_1 = \text{odds}_2$ and $\theta = \text{odds}_1/\text{odds}_2 = 1$. The independence value $\theta = 1$ is a baseline for comparison. Odds ratios on each side of 1 reflect certain types of associations. When $\theta > 1$, the odds of success are higher in row 1 than in row 2. For instance, when $\theta = 4$, the odds of success in row 1 are four times the odds of success in row 2. Thus, subjects in row 1 are more likely to have successes than are subjects in row 2; that is, $\pi_1 > \pi_2$. When $\theta < 1$, a success is less likely in row 1 than in row 2; that is, $\pi_1 < \pi_2$.

Values of θ farther from 1.0 in a given direction represent a stronger association. An odds ratio of 4 is farther from independence than an odds ratio of 2, and an odds ratio of 0.25 is farther from independence than an odds ratio of 0.50. Two values for θ represent the same strength of association, but in opposite directions, when one value is the reciprocal of the other. When $\theta = 0.25$, for example, the odds of success in row 1 are 0.25 times the odds of success in row 2, or equivalently 1/0.25 = 4.0 times as high in row 2 as in row 1. When the order of the rows is reversed or the order of the columns is reversed, the new value of θ is the reciprocal of the original value. This ordering is usually arbitrary, so whether we get 4.0 or 0.25 for the odds ratio is merely a matter of how we label the rows and columns.

The odds ratio does not change value when the table orientation reverses so that the rows become the columns and the columns become the rows. The same value occurs when we treat the columns as the response variable and the rows as the explanatory variable, or the rows as the response variable and the columns as the explanatory variable. It is unnecessary to identify one classification as a response variable in order to estimate θ. In fact, we can also define the odds ratio when *both* variables are response variables, using the joint probabilities,

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

The sample odds ratio equals the ratio of the sample odds in the two rows,

$$\hat{\theta} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{2.2}$$

For a multinomial distribution over the four cells or for independent binomial distributions for the two rows, this is the ML estimator of θ. The odds ratio is also called the *cross-product ratio* because it equals the ratio of the products $n_{11}n_{22}$ and $n_{12}n_{21}$ of cell counts from diagonally opposite cells.

2.3.2 Example: Odds Ratio for Aspirin Use and Heart Attacks

We now find the two odds and the odds ratio for Table 2.2 on aspirin use and myocardial infarction (MI). For the physicians taking placebo, the estimated odds of MI equal $n_{11}/n_{12} = 189/10{,}845 = 0.0174$. Since $0.0174 = 1.74/100$, the value 0.0174 means there were 1.74 *yes* outcomes for every 100 *no* outcomes. For those taking aspirin, the estimated odds equal $104/10{,}933 = 0.0095$, or 0.95 *yes* outcomes per every 100 *no* outcomes.

The sample odds ratio equals $\hat{\theta} = 0.0174/0.0095 = 1.832$. This also equals the cross-product ratio $(189 \times 10{,}933)/(10{,}845 \times 104)$. The estimated odds of MI for those taking placebo equal 1.83 times the estimated odds for those taking aspirin. The estimated odds were 83% higher for the placebo group.

2.3.3 Inference for Odds Ratios and Log Odds Ratios

The sampling distribution of the odds ratio is highly skewed unless the sample size is extremely large. When $\theta = 1$, for example, $\hat{\theta}$ cannot be much smaller than θ (since $\hat{\theta} \geq 0$), but it could be much larger with nonnegligible probability. Because of this skewness, statistical inference uses its natural logarithm, $\log(\theta)$. Independence corresponds to $\log(\theta) = 0$. That is, an odds ratio of 1.0 is equivalent to a log odds ratio of 0.0. The log odds ratio is symmetric about zero, in the sense that reversing rows or reversing columns changes its sign. Two values for $\log(\theta)$ that are the same except for sign, such as $\log(2.0) = 0.69$ and $\log(0.5) = -0.69$, represent the same strength of association.

The sample log odds ratio, $\log\hat{\theta}$, has a less-skewed, bell-shaped sampling distribution. Its approximating normal distribution has a mean of $\log\theta$ and a standard error of

$$SE = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

The SE decreases as the cell counts increase. Because the sampling distribution is closer to normality for $\log\hat{\theta}$ than $\hat{\theta}$, we construct confidence intervals for $\log\theta$ and transform back (that is, take antilogs, using the *exponential function*, discussed below) to form a confidence interval for θ. A large-sample Wald confidence interval for $\log\theta$ is

$$\log\hat{\theta} \pm z_{\alpha/2}(SE).$$

For Table 2.2, the natural log of $\hat{\theta}$ equals $\log(1.832) = 0.605$ and the SE of $\log\hat{\theta}$ is

$$SE = \sqrt{\frac{1}{189} + \frac{1}{10,933} + \frac{1}{104} + \frac{1}{10,845}} = 0.123.$$

A 95% confidence interval for $\log\theta$ equals $0.605 \pm 1.96(0.123)$, or (0.365, 0.846). The corresponding confidence interval for θ is

$$(\exp(0.365), \exp(0.846)) = (e^{0.365}, e^{0.846}) = (1.44, 2.33).$$

The symbol e^x, also expressed as $\exp(x)$, denotes the *exponential function* evaluated at x. The exponential function is the antilog for the logarithm using the natural log scale.[6]

Since the confidence interval (1.44, 2.33) for θ does not contain the "no effect" value of 1.0, the true odds of MI seem different for the two groups. We estimate that the odds of MI are at least 44% higher when taking placebo than when taking aspirin. The endpoints of the interval are not equally distant from $\hat{\theta} = 1.83$, because the sampling distribution of $\hat{\theta}$ is skewed to the right.

For small samples, other methods perform better than the Wald confidence interval, especially when π_1 and π_2 are close to 0 or 1. Computations for the score confidence interval are beyond our scope, but it is available in software. The score confidence interval is available even when an $n_{ij} = 0$, in which case the sample log odds ratio equals $-\infty$ or ∞ and the Wald method fails completely. We next use R to find the Wald and score intervals:

```
------------------------------------------------------------------------------
> library(epitools) # uses the four cell counts
> oddsratio(c(189,10845,104,10933), method="wald", conf=0.95, correct=FALSE)
   odds ratio with 95% C.I.
   estimate   lower    upper
   1.83205  1.44004  2.33078 # Wald CI for odds ratio, no continuity correction

> library(PropCIs) # uses success count and n for each binomial group
> orscoreci(189, 11034, 104, 11037, conf.level=0.95)
95 percent confidence interval:
 1.44080  2.32955  # score CI for odds ratio
------------------------------------------------------------------------------
```

2.3.4 Relationship Between Odds Ratio and Relative Risk

A sample odds ratio of 1.83 does *not* mean that $\hat{\pi}_1$ is 1.83 times $\hat{\pi}_2$, which is the interpretation of a *relative risk* of 1.83, since that measure is a ratio of proportions rather than odds. Instead, $\hat{\theta} = 1.83$ means that the *odds* value $\hat{\pi}_1/(1-\hat{\pi}_1)$ is 1.83 times the odds value $\hat{\pi}_2/(1-\hat{\pi}_2)$.

From (2.2) and the sample value $(\hat{\pi}_1/\hat{\pi}_2)$ of the relative risk,

$$\text{Odds ratio} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \text{Relative risk} \times \left(\frac{1-\hat{\pi}_2}{1-\hat{\pi}_1}\right).$$

When $\hat{\pi}_1$ and $\hat{\pi}_2$ are both close to zero, the fraction in the last term of this expression equals approximately 1.0. The odds ratio and relative risk then take similar values. The example illustrates this similarity. For each group, the sample proportion of MI cases is close to zero. Thus, the sample odds ratio of 1.83 is similar to the sample relative risk of 1.82 that

[6] This means that $e^x = c$ is equivalent to $\log(c) = x$. For instance, $e^0 = 1$ corresponds to $\log(1) = 0$ and $e^{0.69} = 2.0$ corresponds to $\log(2) = 0.69$.

Section 2.2.3 reported. In such a case, an odds ratio of 1.83 *does* mean that $\hat{\pi}_1$ is *approximately* 1.83 times $\hat{\pi}_2$.

This relationship between the odds ratio and the relative risk is useful. For some data sets direct estimation of the relative risk is not possible, yet we can estimate the odds ratio and use it to approximate the relative risk when the proportions compared are small.[7] The next example illustrates this.

2.3.5 Example: The Odds Ratio Applies in Case-Control Studies

Table 2.3 comes from one of the first studies of the association between lung cancer and smoking, based on data from 20 hospitals in London, England, in 1920. At the time, many medical scientists thought that the increased rates of lung cancer in London mainly reflected the increasingly severe air pollution due to the burning of coal (and, thus, the frequent "London fog") before the Clean Air Act of 1956. However, the epidemiologist Richard Doll and statistician Austin Bradford Hill thought that smoking could be a culprit. In their study, patients admitted to the hospital with lung cancer in the preceding year were queried about their smoking behavior. Patients were defined as smokers if they had smoked at least one cigarette a day for at least a year. For each patient admitted, they recorded the smoking behavior of a noncancer patient at the same hospital of the same gender and within the same 5-year grouping on age. The 709 *cases* in the first column of Table 2.3 are those having lung cancer and the 709 *controls* in the second column are those not having it.

Table 2.3 Cross-classification of smoking by lung cancer.

	Lung Cancer	
Smoker	Cases	Controls
Yes	688	650
No	21	59
Total	709	709

Source: Based on data reported in Table IV, R. Doll and A.B. Hill, *British Med. J.* Sept. 30, 1950, 739–748.

Normally, whether lung cancer occurs is a response variable and smoking behavior is an explanatory variable. In this study, however, the marginal distribution (0.50, 0.50) of lung cancer is fixed by the sampling design, and the outcome measured is whether the subject ever was a smoker. The study, which uses a *retrospective* design to "look into the past," is called a *case-control study*. Such studies are common in health-related applications. Often, the two samples are matched, as in this study.

In comparing smokers with nonsmokers on the proportions who suffered lung cancer, the proportions refer to the conditional distribution of lung cancer, given smoking behavior. Instead, case-control studies provide proportions in the reverse direction, for the conditional distribution of smoking behavior, given lung cancer status. For those in Table 2.3 with lung cancer, the proportion who were smokers was $688/709 = 0.970$, while it was $650/709 = 0.917$ for the controls. Because we cannot estimate the conditional

[7] When $\hat{\pi}_1$ and $\hat{\pi}_2$ are above 0.20, the square root of the odds ratio is a rough approximation of the relative risk. See T.J. VanderWeele, *Epidemiology* **28**: 58–60 (2017).

distribution on the response variable, we cannot estimate the difference of proportions or relative risk for the outcome of interest. We can find the odds ratio, however, because it treats the variables symmetrically, taking the same value using the conditional distribution of X given Y as it does using the conditional distribution of Y given X.

For Table 2.3, we can compute the odds ratio using (2.2),

$$\frac{(688/709)/(21/709)}{(650/709)/(59/709)} = \frac{688 \times 59}{650 \times 21} = 3.0.$$

Moreover, by the symmetry, interpretations can use the direction of interest, even though the study was retrospective. The estimated odds of lung cancer for smokers were 3.0 times the estimated odds for nonsmokers.

If the probability of lung cancer is small regardless of smoking behavior, by Section 2.3.4, the odds ratio estimate of 3.0 is also a rough estimate of the relative risk; that is, smokers had about 3.0 times the relative frequency of lung cancer as nonsmokers.[8]

2.3.6 Types of Studies: Observational Versus Experimental

By contrast to the case-control study summarized by Table 2.3, imagine a study that follows a sample of young people for the next 50 years, observing the rates of lung cancer for smokers and nonsmokers. Such a sampling design is *prospective*. Prospective studies are of two types. *Clinical trials* randomly allocate subjects to the two groups of interest, such as in the Physicians' Health Study about aspirin and MI described in Section 2.2.2, observing the response variable in future time. In *cohort studies*, subjects make their own choice about which group to join (e.g., whether to be a smoker) and the study observes the response variable in future time, such as in the Nurses' Health Study.

Yet another approach, a *cross-sectional design*, samples people and classifies them simultaneously on the group classification and their current response. As in a case-control study, we can then gather the data at once, rather than waiting for future events. An example is a sample survey, such as Table 2.1, which cross-classifies gender and belief in an afterlife.

Case-control, cohort, and cross-sectional studies are *observational studies*. We observe who is in each group and who has the outcome of interest. By contrast, a clinical trial is an *experimental study*: the investigator has control, through randomization, over which subjects enter each group, for instance, which subjects take aspirin and which take placebo. Experimental studies have fewer potential pitfalls for comparing groups, because the randomization to groups tends to balance the groups on lurking variables that could be associated both with the response variable and the group identification. Although observational studies are often more practical for biomedical and social science research, they have more potential bias, and it is improper to conclude that an observational association reflects a causal connection.

2.4 CHI-SQUARED TESTS OF INDEPENDENCE

Consider the null hypothesis (H_0) that cell probabilities in a two-way contingency table equal certain fixed values $\{\pi_{ij}\}$. For a sample of size n with cell counts $\{n_{ij}\}$, the values

[8] Recent studies estimate that the lifetime probability of developing lung cancer is >0.2 for heavy smokers and <0.005 for nonsmokers, for an odds ratio in excess of 50.

$\{\mu_{ij} = n\pi_{ij}\}$ are called *expected frequencies*. They represent the expected values $\{E(n_{ij})\}$ when H_0 is true. To judge whether the data contradict H_0, we compare $\{n_{ij}\}$ to $\{\mu_{ij}\}$. If H_0 is true, n_{ij} should be close to μ_{ij} in each cell.

2.4.1 Pearson Statistic and the Chi-Squared Distribution

The *Pearson chi-squared statistic* for testing H_0 is

$$X^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}, \tag{2.3}$$

with the sum taken over all the cells of the table. It was proposed in 1900 by Karl Pearson, the British statistician known also for the Pearson product-moment correlation estimate, among many contributions. For random sampling and randomized experiments, for large sample sizes X^2 has approximately a chi-squared distribution. It takes its minimum value of zero when all $n_{ij} = \mu_{ij}$. For a fixed sample size, greater differences $\{n_{ij} - \mu_{ij}\}$ produce larger X^2 values and stronger evidence against H_0. The P-value is the probability, under H_0, that X^2 is at least as large as the observed value. This is the chi-squared right-tail probability above the observed X^2 value. The chi-squared approximation improves as $\{\mu_{ij}\}$ increase and $\{\mu_{ij} \geq 5\}$ is usually sufficient for a decent approximation.

The chi-squared distribution is indexed by its degrees of freedom (*df*). It is concentrated over nonnegative values, with mean = *df* and standard deviation = $\sqrt{2df}$. As *df* increases, the distribution concentrates around larger values and is more spread out. The distribution is skewed to the right, but it becomes more bell-shaped as *df* increases. Figure 2.2 displays chi-squared distributions having $df = 1$, 5, 10, and 20. The *df* value equals the difference

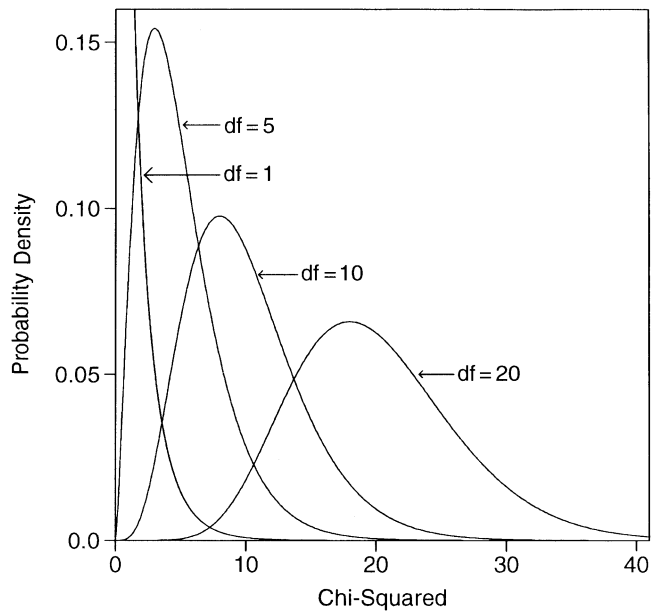


Figure 2.2 Examples of chi-squared distributions.

between the number of parameters in the alternative hypothesis and in the null hypothesis, as we explain in Section 2.4.3.

2.4.2 Likelihood-Ratio Statistic

Of the types of test statistics summarized in Section 1.4.1, the Pearson statistic X^2 is a *score statistic*.[9] An alternative statistic results from the *likelihood-ratio* method for significance tests.

Recall that the likelihood function is the probability of the data, viewed as a function of the parameters, once we observe the data. The likelihood-ratio test determines the parameter values that maximize the likelihood function (a) under the assumption that H_0 is true and (b) under the more general condition that H_0 may or may not be true. As Section 1.4.1 explained, the test statistic uses the ratio of the maximized likelihoods. For two-way contingency tables with likelihood function based on the multinomial distribution, the *likelihood-ratio chi-squared statistic* is

$$G^2 = 2\sum n_{ij} \log\left(\frac{n_{ij}}{\mu_{ij}}\right). \tag{2.4}$$

Like the Pearson statistic, G^2 takes its minimum value of 0 when all $n_{ij} = \mu_{ij}$, and larger values provide stronger evidence against H_0.

The Pearson X^2 and likelihood-ratio G^2 provide separate test statistics, but they share many properties and usually provide the same conclusions. When H_0 is true and the expected frequencies are large, the two statistics have the same chi-squared distribution and their numerical values are similar.

2.4.3 Testing Independence in Two-Way Contingency Tables

In two-way contingency tables with joint probabilities $\{\pi_{ij}\}$ for two response variables, the null hypothesis of statistical independence is

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \quad \text{for all } i \text{ and } j.$$

To test H_0, we identify $\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ as the expected frequency of n_{ij}, assuming independence. Usually, $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are unknown, as is this expected value. To obtain *estimated expected frequencies*, we substitute sample proportions for the unknown marginal probabilities, giving

$$\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \frac{n_{i+}n_{+j}}{n}$$

based on the row marginal totals $\{n_{i+}\}$ and the column marginal totals $\{n_{+j}\}$. The $\{\hat{\mu}_{ij}\}$ have the same row and column totals as the cell counts $\{n_{ij}\}$, but they display the pattern of independence.

For testing independence in $r \times c$ contingency tables, the approximate chi-squared sampling distributions of X^2 and G^2 have $df = (r-1)(c-1)$. The df value means the following: under H_0, $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ determines the cell probabilities. There are $r-1$

[9] This means that X^2 is derived using a covariance matrix for $\{n_{ij}\}$ estimated under H_0.

nonredundant row probabilities. Because they sum to 1, the first $r-1$ determines the last one through $\pi_{r+} = 1 - (\pi_{1+} + \cdots + \pi_{r-1,+})$. Similarly, there are $c-1$ nonredundant column probabilities, so, under H_0, there are $(r-1)+(c-1)$ parameters. The alternative hypothesis H_a states that there is not independence but does not specify a pattern for the rc cell probabilities. The probabilities are then solely constrained to sum to 1, so there are $rc-1$ nonredundant parameters. The value for df is the difference between the number of parameters under H_a and H_0, or

$$df = (rc-1) - [(r-1)+(c-1)] = rc - r - c + 1 = (r-1)(c-1).$$

2.4.4 Example: Gender Gap in Political Party Affiliation

Table 2.4, from the 2016 General Social Survey, cross-classifies gender and political party identification (ID). Subjects indicated whether they identified more strongly with the Democratic or Republican party or as Independents. Table 2.4 also contains estimated expected frequencies for H_0: independence. For instance, the first cell has $\hat{\mu}_{11} = n_{1+}n_{+1}/n = (1357 \times 825)/2450 = 456.9$.

Table 2.4 Political party identification by gender, with estimated expected frequencies for independence in parentheses.

	Political Party Identification			
Gender	Democrat	Republican	Independent	Total
Female	495 (456.9)	272 (297.4)	590 (602.6)	1357
Male	330 (368.1)	265 (239.6)	498 (485.4)	1093
Total	825	1088	2450	

The chi-squared test statistics are $X^2 = 12.57$ and $G^2 = 12.60$, with $df = (r-1)(c-1) = (2-1)(3-1) = 2$. This chi-squared distribution has a mean of $df = 2$ and a standard deviation of $\sqrt{2df} = \sqrt{4} = 2$. Therefore, a value of 12.6 is well out in the right-hand tail. Each statistic has a P-value of 0.002. This evidence of association would be rather unusual if the variables were truly independent. Both test statistics suggest that political party ID and gender are associated.

2.4.5 Residuals for Cells in a Contingency Table

A test statistic and its P-value describe the evidence against H_0. A cell-by-cell comparison of observed and estimated expected frequencies helps us better understand the nature of the evidence. Larger differences between n_{ij} and $\hat{\mu}_{ij}$ tend to occur for cells that have larger expected frequencies, so the raw difference $n_{ij} - \hat{\mu}_{ij}$ is insufficient. For the test of independence, a more useful cell residual is

$$\frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-\hat{\pi}_{i+})(1-\hat{\pi}_{+j})}} = \frac{n(\hat{\pi}_{ij} - \hat{\pi}_{i+}\hat{\pi}_{+j})}{\sqrt{n\hat{\pi}_{i+}\hat{\pi}_{+j}(1-\hat{\pi}_{i+})(1-\hat{\pi}_{+j})}}. \tag{2.5}$$

The denominator is the estimated standard error of the numerator, under H_0. This measure is called a *standardized residual*, because of this standardization.

When H_0 is true, each standardized residual has a large-sample standard normal distribution. A standardized residual having an absolute value that exceeds about 2 when there

are few cells or about 3 when there are many cells indicates lack of fit of H_0 in that cell. (Under H_0, we expect about 5% of the standardized residuals to be farther from 0 than ± 2 by chance alone.)

Table 2.5 shows the standardized residuals for testing H_0: independence with Table 2.4. For the first cell, for instance, the standardized residual of 3.27 indicates that $n_{11} - \hat{\mu}_{11}$ is 3.27 standard errors from 0, a greater discrepancy than we would be expected if the variables were truly independent. Large *positive* residuals occur for female Democrats and male Republicans: more female Democrats and male Republicans occurred than the hypothesis of independence predicts. Large *negative* residuals occur for female Republicans and male Democrats. Fewer females were Republicans and fewer males were Democrats than the hypothesis of independence predicts.

Table 2.5 Observed frequencies for political party identification and gender, with standardized residuals in parentheses for test of independence.

	Political Party Identification		
Gender	Democrat	Republican	Independent
Female	495 (3.27)	272 (−2.50)	590 (−1.03)
Male	330 (−3.27)	265 (2.50)	498 (1.03)

For each political party, Table 2.5 shows that the residual for females is the negative of the one for males. This is because the observed counts and the estimated expected frequencies have the same row and column totals. In a particular column, if $n_{ij} > \hat{\mu}_{ij}$ in one cell, the reverse must happen in the other cell. The differences $n_{1j} - \hat{\mu}_{1j}$ and $n_{2j} - \hat{\mu}_{2j}$ have the same magnitude but different signs, implying the same pattern for their standardized residuals.

Odds ratios describe the association. The 2×2 table of Democrat and Republican identifiers has a sample odds ratio of $(495 \times 265)/(272 \times 330) = 1.46$. Of those subjects identifying with one of the two parties, the estimated odds of identifying with the Democrats rather than the Republicans were 46% higher for females than males.

In R, with the data file first read from the text website and then tabulated in a contingency table, here is how you can perform the Pearson chi-squared test of independence and generate standardized residuals:

```
-------------------------------------------------------------------------------
> Political <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Political.dat",
+                         header=TRUE)
> Political
person gender party # data file has 2450 lines, one for each person
     1 female   Dem # enter head(Political) to show just first 6 lines
     2 female   Dem
...
  2450   male   Ind
> Party <- factor(Political$party, levels = c("Dem","Rep","Ind"))
> # levels specifies preferred order of categories for displays
> GenderGap <- xtabs(~gender + Party, data=Political) # forms contingency table
> GenderGap
        Party
gender   Dem Rep Ind
```

```
  female 495 272 590
  male   330 265 498
> # If you already know the continency table counts, you can read in data by
  # GenderGap <- matrix(c(495,272,590,330,265,498), ncol=3, byrow=TRUE)
> chisq.test(GenderGap)
        Pearson's Chi-squared test
X-squared = 12.569, df = 2, p-value = 0.001865
> stdres <- chisq.test(GenderGap)$stdres # standardized residuals
> stdres
       Party
gender          Dem       Rep       Ind
  female   3.27236  -2.49856  -1.03220
  male    -3.27236   2.49856   1.03220

> library(vcd) # useful package for graphics, beyond scope of this book
> mosaic(GenderGap, gp=shading_Friendly, residuals=stdres,      # mosaic plot
+       residuals_type="Std\nresiduals", labeling=labeling_residuals)
-------------------------------------------------------------------------------
```

The end of this code has a function to create a *mosaic plot*, shown in Figure 2.3. It graphically portrays the cell counts and the standardized residuals. For this and more detailed graphical techniques for categorical data using R, see Friendly and Meyer (2016) and the vcd, vcdExtra, and extracat packages.

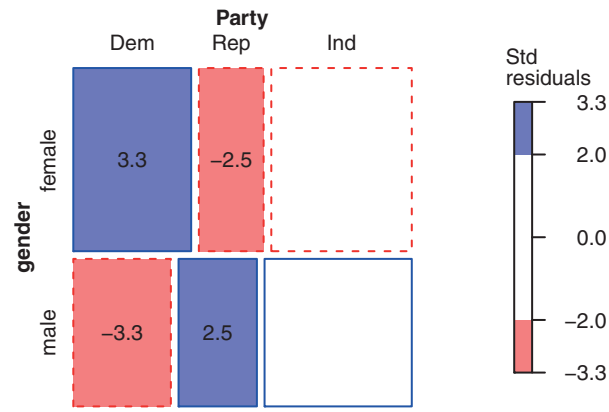


Figure 2.3 Mosaic plot portraying relative sizes of cell counts and of standardized residuals for Table 2.4.

2.4.6 Partitioning Chi-Squared Statistics

Chi-squared statistics sum to and decompose into other chi-squared statistics. If a chi-squared statistic has $df = df_1$ and a separate, independent, chi-squared statistic has $df = df_2$, then their sum has a chi-squared distribution with $df = df_1 + df_2$. Likewise, chi-squared statistics having $df > 1$ can be broken into components with fewer degrees of freedom. Another supplement to a test of independence partitions its chi-squared test statistic so that the components represent certain aspects of the association. A partitioning

may show that an association primarily reflects differences between certain categories or groupings of categories.

To illustrate, the G^2 statistic for testing independence in Table 2.4, which has size 2×3, has $df = 2$. This chi-squared statistic can partition into two components. For example, $G^2 = 12.60$ equals the sum of a G^2 statistic that compares the first two columns, plus a G^2 statistic for the 2×2 table that combines the first two columns and compares them to the third column. Each component G^2 statistic has $df = 1$. The first two columns form a 2×2 table with cell counts, by row, of (495, 272 / 330, 265). For this table, $G^2 = 11.536$, with $df = 1$. Of those subjects who identify either as Democrats or Republicans, there is strong evidence (P-value < 0.001) of a difference between females and males in the relative numbers in the two categories. The second 2×2 table combines these columns and compares them to the Independent column, giving the table with rows $(495 + 272, 590 / 330 + 265, 498) = (767, 590 / 595, 498)$. This table has $G^2 = 1.065$, based on $df = 1$. There is no evidence of a difference between females and males in the relative numbers identifying as Independent instead of Democrat or Republican.

Note that $11.536 + 1.065 = 12.60$; that is, the sum of these G^2 components equals G^2 for the test of independence for the full 2×3 table. The Pearson X^2 for the full table does not equal the sum of X^2 values for the separate tables. It is valid to use the X^2 statistics for those tables, but they do not provide an exact algebraic partitioning of X^2.

It might seem more natural to find G^2 for separate 2×2 tables that pair each column with a particular one, say the last. This is a reasonable way to investigate association in many data sets. However, these component statistics are not independent and do not sum exactly to G^2 for the complete table.

2.4.7 Limitations of Chi-Squared Tests

Chi-squared tests of independence, like any significance test, have limitations. They merely indicate the degree of evidence for an association. They are rarely adequate for answering all questions that a study poses. Rather than relying solely on these tests, it is sensible to analyze the nature of the association by finding standardized residuals and estimating parameters such as odds ratios that describe the strength of association.

The chi-squared tests also have limitations in the types of data sets for which they are applicable. For instance, they require large samples. The sampling distributions of X^2 and G^2 get closer to chi-squared as n increases. The convergence is quicker for X^2 than G^2. The chi-squared approximation is often poor for G^2 when some expected frequencies are less than about 5. To be cautious, you should instead use a small-sample test whenever at least one expected frequency is less than 5. Section 2.6 presents small-sample methods. Although computationally more intensive, these less-restrictive tests are now available in software for any n, not merely small values.

The values of the chi-squared statistics do not depend on the order in which the rows and columns are listed. Thus, these tests treat both classifications as nominal scale, that is, having unordered categories. When at least one variable is ordinal, more powerful tests of independence usually exist. The next section presents such tests.

2.5 TESTING INDEPENDENCE FOR ORDINAL VARIABLES

An ordinal variable has naturally ordered categories. An example is a rating of customer service using categories (excellent, good, fair, poor). When the rows and/or the columns

of a contingency table are ordinal, the chi-squared test of independence using test statistic X^2 or G^2 ignores the ordering information. Test statistics can use the ordinality by treating ordinal variables in a quantitative manner.

2.5.1 Linear Trend Alternative to Independence

When the variables are ordinal, a trend association is common. As the level of X increases, responses on Y tend to increase toward higher levels, or responses on Y tend to decrease toward lower levels. To detect a trend association, a simple analysis assigns ordered scores to categories and measures the degree of *linear trend*. The test statistic, which is sensitive to positive or negative linear trends, utilizes correlation information in the data. Let $u_1 \leq u_2 \leq \cdots \leq u_r$ denote scores for the rows and $v_1 \leq v_2 \leq \cdots \leq v_c$ denote scores for the columns, having the same ordering as the categories. You could choose the scores to reflect distances between categories, with greater distances between categories regarded as farther apart.

Let $\bar{u} = \sum_i u_i \hat{\pi}_{i+}$ denote the sample mean of the row scores and let $\bar{v} = \sum_j v_j \hat{\pi}_{+j}$ denote the sample mean of the column scores. The sample correlation R between X and Y is

$$R = \frac{\sum_{i,j}(u_i - \bar{u})(v_j - \bar{v})\hat{\pi}_{ij}}{\sqrt{\left[\sum_i (u_i - \bar{u})^2 \hat{\pi}_{i+}\right]\left[\sum_j (v_j - \bar{v})^2 \hat{\pi}_{+j}\right]}}.$$

The correlation falls between -1 and $+1$. Independence between the variables implies that its population value $\rho = 0$. The larger the value of $|R|$, the farther the data fall from independence in the linear dimension.

For testing H_0: independence against H_a: $\rho \neq 0$, the test statistic

$$M^2 = (n-1)R^2 \tag{2.6}$$

is a special case of a statistic for stratified ordinal contingency tables.[10] This test statistic increases as $|R|$ increases and as the sample size n grows. For large n, M^2 has approximately a chi-squared distribution with $df = 1$. Large values contradict independence, so, as with X^2 and G^2, the P-value is the right-tail probability above the observed value. The square root, $M = \sqrt{n-1}R$, has approximately a standard normal null distribution. It applies to one-sided alternatives, such as H_a: $\rho > 0$.

Like X^2 and G^2, M^2 does not distinguish between response and explanatory variables. We get the same value regardless of which is the row variable and which is the column variable.

2.5.2 Example: Alcohol Use and Infant Malformation

Table 2.6 comes from a prospective study of the effects of maternal drinking. After the first three months of pregnancy, the women in the sample completed a questionnaire about alcohol consumption. Following childbirth, observations were recorded on the presence or

[10] A score test, proposed by Nathan Mantel in 1963.

Table 2.6 Infant malformation and mother's alcohol consumption.

Alcohol	Malformation			Percentage	Standardized
Consumption	Absent	Present	Total	Present	Residual
0	17,066	48	17,114	0.28	−0.18
< 1	14,464	38	14,502	0.26	−0.71
1–2	788	5	793	0.63	1.84
3–5	126	1	127	0.79	1.06
≥ 6	37	1	38	2.63	2.71

Source: Graubard, B.I., and Korn, E.L., *Biometrics* **43**: 471–476 (1987); reprinted with permission from the Biometric Society.

absence of congenital sex organ malformations. Alcohol consumption, measured as the average number of drinks per day, is an explanatory variable with ordered categories. Malformation, the response variable, is binary.

When a variable is binary, statistics such as M^2 that treat the variable as ordinal are still valid. For instance, we could artificially regard malformation as ordinal, treating *absent* as *low* and *present* as *high.* Any choice of two scores, such as 0 for *absent* and 1 for *present,* yields the same value of M^2.

Table 2.6 has a mixture of very small, moderate, and extremely large counts. Even though the sample size is large ($n = 32{,}574$), in such cases the actual sampling distributions of X^2 or G^2 may not be close to chi-squared. For these data, having $df = 4$, $G^2 = 6.2$ ($P = 0.19$) and $X^2 = 12.1$ ($P = 0.02$) provide mixed signals. In any case, they ignore the ordinality of alcohol consumption.

Table 2.6 also reports the sample percentage of malformation cases and the standardized residuals for the *present* category. These both suggest a possible tendency for malformations to be more likely at higher levels of alcohol consumption. To use the ordinal test statistic M^2, we assign scores to alcohol consumption that are midpoints of the categories; that is, $v_1 = 0$, $v_2 = 0.5$, $v_3 = 1.5$, $v_4 = 4.0$, $v_5 = 7.0$, the last score being somewhat arbitrary. The sample correlation between alcohol consumption and malformation is $R = 0.0142$. This seems weak, but for tables for which a binary variable has greatly different marginal totals, it is not possible to obtain a large value for R. The value of R is not useful here for describing strength of association, but it is sufficient for summarizing trend information for the test. The test statistic $M^2 = (n-1)R^2 = (32573)(0.0142)^2 = 6.57$ has P-value = 0.010, suggesting strong evidence of a nonzero correlation. The standard normal statistic $M = 2.56$ has $P = 0.005$ for H_a: $\rho > 0$. Here is R code for the test:

```
------------------------------------------------------------------------------
> Malform <- matrix(c(17066, 14464, 788, 126, 37, 48, 38, 5, 1, 1), ncol=2)
> Malform
      [,1] [,2]
[1,] 17066   48
[2,] 14464   38
[3,]   788    5
[4,]   126    1
[5,]    37    1
> library(vcdExtra)
```

```
> CMHtest(Malform, rscores = c(0, 0.5, 1.5, 4.0, 7.0)) # row scores
Cochran-Mantel-Haenszel Statistics # corr. case was proposed by Nathan Mantel
                 AltHypothesis   Chisq  Df     Prob
cor        Nonzero correlation  6.5699   1  0.01037 # M-squared = 6.5699
> sqrt(6.5699) # M test statistic
[1] 2.56318
> 1 - pnorm(2.56318) # one-sided standard normal P-value for M statistic
[1] 0.00519
```

Future chapters present tests such as M^2 as part of a model-based analysis that yields estimates of the effect size as well as smoothed estimates of cell probabilities. These estimates are more informative than mere significance tests.

2.5.3 Ordinal Tests Usually Have Greater Power

For testing H_0: independence, X^2 and G^2 refer to the most general H_a possible, whereby cell probabilities exhibit *any* pattern for the statistical dependence. Their df value of $(r-1)(c-1)$ reflects that H_a has $(r-1)(c-1)$ more parameters than H_0 (recall the discussion at the end of Section 2.4.3). In achieving this generality, the statistics sacrifice sensitivity for detecting particular patterns.

The M^2 test statistic describes the association using a single extra parameter, based on a correlation measure of linear trend. When a chi-squared test statistic refers to a single parameter, it has $df = 1$. When the association truly has a positive or negative trend, the ordinal test using M^2 has a power advantage over the tests based on X^2 or G^2. Since df equals the mean of the chi-squared distribution, a relatively large M^2 value based on $df = 1$ falls farther out in its right-hand tail than a comparable value of X^2 or G^2 based on $df = (r-1)(c-1)$. Falling farther out in the tail produces a smaller P-value. When there truly is a linear trend, M^2 often has a similar size as X^2 or G^2, so it tends to provide smaller P-values.

2.5.4 Choice of Scores

For most data sets, different choices of ordered scores yield similar test results. This may not happen, however, when the data are very unbalanced, such as when some categories have many more observations than other categories. Table 2.6 illustrates this. For the equally spaced row scores (1, 2, 3, 4, 5), $M^2 = 1.83$, giving a much weaker conclusion ($P = 0.18$). The magnitudes of the correlation R and M^2 do not change with transformations of the scores that maintain the same relative spacings between the categories. For example, scores (1, 2, 3, 4, 5) yield the same correlation as scores (0, 1, 2, 3, 4) or (2, 4, 6, 8, 10) or (10, 20, 30, 40, 50).

An alternative approach assigns ranks to the subjects and uses them as the category scores, but usually it is better to use your judgment by selecting scores that reflect well the distances between categories. When uncertain, perform a sensitivity analysis. Select two or three sensible choices and check that the results are similar for each. Equally spaced scores often are a reasonable compromise when the category labels do not suggest any obvious choices, such as the categories (liberal, moderate, conservative) for political ideology.

Alternative ordinal tests for $r \times c$ tables utilize versions of ordinal association measures that use the ordering information without assigning scores. For instance, *gamma* and *Kendall's tau-b* are contingency-table generalizations of the nonparametric statistic called *Kendall's tau*. The sample value of any such measure divided by its standard error has a large-sample standard normal distribution for testing independence. Like the test based on M^2, these tests share the potential power advantage that results from using a single parameter to describe the association.

2.5.5 Trend Tests for $r \times 2$ and $2 \times c$ and Nominal–Ordinal Tables

When Y is binary with ordinal X, the table has size $r \times 2$, such as Table 2.6. We then focus on how the proportion of successes varies across the levels of X. For the chosen row scores, M^2 detects a linear trend in this proportion and relates to models we shall present in Section 3.2.1. Small P-values suggest that the population slope for this linear trend is nonzero. This special case of the ordinal test[11] is called the *Cochran–Armitage trend test*.

When X is binary, the table has size $2 \times c$. Such tables occur in comparisons of two groups, such as when the rows represent two treatments. The M^2 statistic then detects differences between the two row means of the scores $\{v_j\}$ on Y. Small P-values suggest that the true difference in row means is nonzero.

When X is nominal-scale with $r > 2$ categories and Y is ordinal, the M^2 test is invalid because it treats both classifications as ordinal. Such a case can be dealt with by a model for an ordinal response variable that we shall present in Section 6.2 that can have a nominal-scale explanatory variable. Statistics for testing independence then have a large-sample chi-squared distribution with $df = (r - 1)$.

2.6 EXACT FREQUENTIST AND BAYESIAN INFERENCE *

The confidence intervals and tests presented so far in this chapter are large-sample methods. As the sample size n grows in a randomized study, *chi-squared* statistics such as Pearson's X^2 have sampling distributions that are more nearly chi-squared. When n is small or when the data are so unbalanced that one suspects possible irregularities in the sampling distribution, frequentist inference can use *exact* distributions rather than large-sample approximations. In fact, computational methods are now sufficiently developed that you can very precisely approximate exact distributions with *any* sample size. Therefore, you can *always* use exact tests instead of chi-squared tests for testing independence. For any n, exact inference is also available using the Bayesian approach.

2.6.1 Fisher's Exact Test for 2×2 Tables

For 2×2 tables, independence corresponds to an odds ratio of $\theta = 1$. Suppose the cell counts $\{n_{ij}\}$ result from two independent binomial samples or from a single multinomial sample over the four cells. A small-sample null probability distribution for the cell counts that does not depend on any unknown parameters results from considering the set of

[11] Proposed by Peter Armitage in 1955 and William Cochran in 1954.

tables having the same row and column totals as the observed data. When the cell counts result from two independent binomial samples, the row totals (i.e., the binomial n indices) are already fixed and we consider the tables that also have the same column totals as the observed data. Once we condition on this restricted set of tables, the cell counts have exactly the *hypergeometric* distribution for any n, with no parameters.

For fixed row and column marginal totals, n_{11} determines the other three cell counts. Thus, the hypergeometric formula expresses probabilities for the four cell counts in terms of n_{11} alone. When $\theta = 1$, the probability of a particular value n_{11} equals

$$P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}}, \quad \text{where} \quad \binom{a}{b} = \frac{a!}{b!(a-b)!}. \tag{2.7}$$

This probability distribution does not have any unknown parameters, so it allows exact rather than approximate P-value calculations.

To test H_0: independence, the P-value is the sum of hypergeometric probabilities for outcomes at least as favorable to H_a as the observed outcome. For H_a: $\theta > 1$, given the marginal totals, tables having larger n_{11} values also have larger sample odds ratios $\hat{\theta} = (n_{11}n_{22})/(n_{12}n_{21})$; hence, they provide stronger evidence in favor of this alternative. The P-value equals the right-tail hypergeometric probability that n_{11} is at least as large as the observed value. This test, proposed by the eminent British statistician R.A. Fisher in 1934, is called *Fisher's exact test*.

2.6.2 Example: Fisher's Tea Tasting Colleague

To illustrate this test in his 1935 book, *The Design of Experiments*, Fisher described the following experiment. When drinking tea, Dr. Muriel Bristol, a colleague of Fisher's at Rothamsted Experiment Station near London, claimed she could distinguish whether milk or tea was added to the cup first. To test her claim, Fisher designed an experiment in which she tasted eight cups of tea. Four cups had milk added first and the other four had tea added first. She was told there were four cups of each type and she should try to select the four that had milk added first. The cups were presented to her in random order.

Table 2.7 shows a potential result of the experiment. The null hypothesis H_0: $\theta = 1$ for Fisher's exact test states that her guess was independent of the actual order of pouring. The alternative hypothesis that reflects her claim, predicting a positive association between true order of pouring and her guess, is H_a: $\theta > 1$. For this experimental design, the column

Table 2.7 Potential result for Fisher's tea tasting experiment.

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

margins are identical to the row margins (4, 4), because she knew that four cups had milk added first. Both marginal distributions are naturally fixed.

The null distribution of n_{11} is the hypergeometric distribution defined for all 2×2 tables having row and column margins (4, 4). The potential values for n_{11} are (0, 1, 2, 3, 4). For Table 2.7, three correct guesses of the four cups having milk added first, the probability (2.7) is

$$P(3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{\left[\frac{4!}{(3!)(1!)}\right]\left[\frac{4!}{(1!)(3!)}\right]}{\left[\frac{8!}{(4!)(4!)}\right]} = \frac{16}{70} = 0.229.$$

For H_a: $\theta > 1$, the only table that is more extreme in the direction of H_a consists of four correct guesses. It has $n_{11} = n_{22} = 4$ and $n_{12} = n_{21} = 0$, and a probability of

$$P(4) = \binom{4}{4}\binom{4}{0} \Big/ \binom{8}{4} = \frac{1}{70} = 0.014.$$

Table 2.8 summarizes the possible values of n_{11} and their hypergeometric probabilities. This table also shows that the Pearson X^2 statistic has sampling distribution far from chi-squared; conditional on the marginal counts, it can take only the values 0, 2, and 8.

Table 2.8 Hypergeometric distribution for Table 2.7, hypergeometric P-value for H_a: $\theta > 1$, and Pearson's X^2 statistic.

n_{11}	Probability	P-value	X^2
0	0.014	1.000	8.0
1	0.229	0.986	2.0
2	0.514	0.757	0.0
3	0.229	0.243	2.0
4	0.014	0.014	8.0

The P-value for H_a: $\theta > 1$ equals the right-tail hypergeometric probability that n_{11} is at least as large as observed. With $n_{11} = 3$, this P-value is $P(3) + P(4) = 0.243$, not much evidence against H_0: independence. The experiment did not establish an association between the actual order of pouring and the guess, but it is difficult to show effects with such a small sample. For the potential n_{11} values, Table 2.8 shows the possible P-values for H_a: $\theta > 1$. If the tea taster had guessed all cups correctly (i.e., $n_{11} = 4$), the observed result would have been the most extreme possible in the right tail of the hypergeometric distribution. Then, the P-value is $P(4) = 0.014$, giving more reason to believe her claim.[12]

The two-sided H_a: $\theta \neq 1$ is the general alternative hypothesis of statistical dependence, as in chi-squared tests. Its exact P-value is the two-tailed sum of the probabilities of tables no more likely than the observed table. For Table 2.7, summing all probabilities that are

[12] Results of the actual experiment were not recorded, but Joan Fisher Box (a daughter of Fisher) told me that Dr. Bristol did convince Fisher she could detect the actual order.

no greater than the probability $P(3) = 0.229$ of the observed table gives P-value equal to $P(0) + P(1) + P(3) + P(4) = 0.486$. When the row or column marginal totals are equal, the hypergeometric distribution is unimodal and symmetric and the two-sided P-value doubles the one-sided one.

Here is edited R output with two-sided and one-sided Fisher's exact test:

```
-------------------------------------------------------------------------
> tea <- matrix(c(3,1,1,3), ncol=2)
> fisher.test(tea)
Fisher's Exact Test for Count Data
p-value = 0.4857 # two-sided
alternative hypothesis: true odds ratio is not equal to 1

> fisher.test(tea, alternative="greater")
Fisher's Exact Test for Count Data
p-value = 0.2429 # one-sided
alternative hypothesis: true odds ratio is greater than 1
-------------------------------------------------------------------------
```

Exact tests of independence for tables of size larger than 2×2 use a multivariate version of the hypergeometric distribution. Such tests can be computationally infeasible when n is large, but software can use Monte Carlo simulation[13] to approximate very closely the exact P-value. For example, Table 2.4 on political party ID and gender has exact P-value $=$ 0.00185, essentially the same as the P-value of 0.00186 for $X^2 = 12.569$ with the large-sample Pearson chi-squared test.

2.6.3 Conservatism for Actual *P*(Type I Error); Mid *P*-Values

For small samples, the exact distribution (2.7) has relatively few possible values for n_{11} and for the P-value. As Section 1.4.3 explained, discreteness affects error rates. Fisher's exact test is *conservative*, because the actual probability of a Type I error is smaller than a nominal value such as 0.05.

To diminish the conservativeness, we recommend using the *mid P-value*. From Section 1.4.3, this is *half* the probability of the observed result plus the probability of more extreme results. For the tea-tasting data, with $n_{11} = 3$, the one-sided mid P-value equals $P(3)/2 + P(4) = 0.229/2 + 0.014 = 0.129$, compared to 0.243 for the ordinary P-value. In R, here[14] are one-sided and two-sided mid P-values:

```
-------------------------------------------------------------------------
> library(epitools)
> ormidp.test(3, 1, 1, 3, or=1) # enter the four cell counts and H0 value
  one.sided  two.sided
  0.12857    0.25714  # mid P-values for testing independence
-------------------------------------------------------------------------
```

[13] In R, with the `simulate.p.value` option in the `fisher.test` command, an approach that was proposed by A. Agresti, D. Wackerly, and J. Boyett in *Psychometrika* **44**: 75–83 (1979).

[14] Exact inferences for 2×2 tables are also available in the `exact2x2` package.

2.6.4 Small-Sample Confidence Intervals for Odds Ratio

Confidence intervals for the odds ratio can also be based on exact distributions. They correspond to a generalization of Fisher's exact test that tests an arbitrary value, H_0: $\theta = \theta_0$. A 95% confidence interval contains all θ_0 values for which an exact test of H_0: $\theta = \theta_0$ has P-value > 0.05.

As happens with exact tests, discreteness makes these confidence intervals conservative. The true confidence level may be considerably larger than the nominal one, such as 0.95. To reduce the conservativeness, we recommend constructing the confidence interval that corresponds to the test using a mid P-value. This interval is shorter and has actual coverage probability that tends to be closer to the nominal level. For the tea-tasting data (Table 2.7), the 95% confidence interval based on the test using the mid-P value equals (0.31, 306.6), as shown in the next R output:

```
------------------------------------------------------------------------------
> library(epitools)
> or.midp(c(3, 1, 1, 3), conf.level=0.95)$conf.int
   0.31005  306.63385 # mid-P confidence interval for odds ratio
------------------------------------------------------------------------------
```

The interval is very wide, because the sample size is so small.

Alternative exact inferential methods are now available for two independent binomial samples that do not require conditioning on the other margin. These are available for tests and confidence intervals.[15] Beyond the scope of this book, details are in Agresti (2013, Sections 3.5, 16.5, 16.6).

2.6.5 Bayesian Estimation for Association Measures

Bayesian methods are straightforward for estimating association measures for contingency tables. We illustrate for the comparison of parameters for two independent binomial samples summarized in a 2×2 contingency table. We assume that the number of successes Y_1 in row 1 has a binomial distribution for n_1 trials and parameter π_1 and the number of successes Y_2 in row 2 has a binomial distribution for n_2 trials and parameter π_2.

The conjugate Bayesian approach uses independent beta prior distributions, beta(α_1, β_1) for π_1 and beta(α_2, β_2) for π_2, most commonly with all hyperparameters equal to 1 (the *uniform distribution*) or 0.50 (the *Jeffreys prior*). The beta choice of prior distributions yields independent posterior beta($y_i + \alpha_i, n_i - y_i + \beta_i$) distributions for π_i, $i = 1, 2$, which induce corresponding posterior distributions for the difference of proportions, relative risk, and odds ratio. Software can use these to construct posterior intervals for these association measures. To obtain adequate frequentist performance in terms of maintaining coverage probability close to the nominal level over the entire parameter space, a good default choice is the Jeffreys prior.

A caveat: Some software can find an alternative posterior interval, the *highest posterior density* (HPD) interval. It has a higher posterior density for every value inside the interval

[15] For example, for an odds ratio, the method proposed in the article by A. Agresti and Y. Min, *Biostatistics* **3**: 379–386 (2002) is available with the `uncondExact2x2` function in the `exact2x2` package.

than for every value outside it. This method produces the shortest possible interval with the given confidence level. However, it is usually inappropriate for nonlinear functions of probabilities such as the odds ratio and relative risk, because the HPD interval for a random variable X is not the inverse mapping of the HPD interval for a nonlinear transform of it, such as $1/X$. For example, suppose (2.0, 3.0) is the 95% HPD interval for the odds ratio θ. Then, the 95% HPD interval based on the posterior distribution of $1/\theta$, which is relevant if we reverse the labeling of the two groups being compared or of *success* and *failure*, is not $(1/3, 1/2)$. We recommend instead the equal-tail posterior interval.[16]

2.6.6 Example: Bayesian Inference in a Small Clinical Trial

In many biomedical studies for treating a disease, one group (say, group 1) receives a new treatment and the other group receives the standard treatment or placebo, and the study analyzes whether the response tends to be better with the new treatment. When we regard $\pi_1 \leq \pi_2$ as a null condition and $\pi_1 > \pi_2$ as an alternative, the posterior $P(\pi_1 \leq \pi_2)$ is a sort of Bayesian P-value.

With small samples, results can depend strongly on the choice of prior distribution. We illustrate this by an example from a clinical trial[17] that used an urn sampling method to allocate patients to treatments. The 11 patients allocated to the experimental treatment were all successes (i.e., $n_1 = y_1 = 11$) and the only patient allocated to the control treatment was a failure (i.e., $n_2 = 1$, $y_2 = 0$). That is, the 2×2 table has first-row counts (11, 0) for the experimental treatment and second-row counts (0, 1) for the control. With beta(0.5, 0.5) prior distributions for π_1 and π_2, the posterior distribution is beta$(y_1 + 0.5, n_1 - y_1 + 0.5)$ = beta(11.5, 0.5) for π_1 and beta$(y_2 + 0.5, n_2 - y_2 + 0.5)$ = beta(0.5, 1.5) for π_2. Software reports that the 95% equal-tail posterior interval for the odds ratio is (3.3, 1,361,274). By contrast, if we use uniform priors, the 95% posterior interval is (1.7, 4677). However, with such a small n, different frequentist methods can also give quite different results; for example, the 95% confidence interval for the odds ratio is $(4.5, \infty)$ using the large-sample score method and $(1.22, \infty)$ using the small-sample mid P-value method.

Here are very precise simulation results for 95% equal-tail posterior intervals for this example, using beta(0.5, 0.5) prior distributions:

```
------------------------------------------------------------------------------
> library(PropCIs)
> orci.bayes(11, 11, 0, 1, 0.5, 0.5, 0.5, 0.5, 0.95, nsim = 1000000)
> # arguments are y1, n1, y2, n2, alpha1, beta1, alpha2, beta2, post. prob.
[1] 3.276438e+00  1.361274e+06  # posterior interval for odds ratio

> diffci.bayes(11, 11, 0, 1, 0.5, 0.5, 0.5, 0.5, 0.95, nsim = 1000000)
[1] 0.09900  0.99327  # posterior interval for difference of proportions
------------------------------------------------------------------------------
```

A simple but less precise way to approximate posterior intervals for measures such as the odds ratio and difference of proportions uses direct simulation: generate a very large

[16] For discussion of Bayesian methods for 2×2 tables, see the article by A. Agresti and Y. Min, *Biometrics* **61**: 515–523 (2005).

[17] Example from the article by C.B. Begg in *Biometrika* **77**: 467–484 (1990).

number of beta random variables from the posterior beta densities of π_1 and π_2, calculate the measure for each generate, and find quantiles corresponding to desired probabilities. You can also approximate the posterior $P(\pi_1 \leq \pi_2)$ by the proportion of simulated cases for which $\pi_1 \leq \pi_2$. We again use beta(0.5, 0.5) priors and the corresponding beta(11.5, 0.5) and beta(0.5, 1.5) posterior distributions for π_1 and π_2:

```
-------------------------------------------------------------------------------
> pi1 = rbeta(100000000, 11.5, 0.5); pi2 = rbeta(100000000, 0.5, 1.5)
   # simulate 100000000 posterior beta(11.5, 0.5) and beta(0.5, 1.5)
   # which result from beta(0.5, 0.5) prior distributions
   # use huge number of simulations to approx. interval for odds ratio well
> or <- pi1*(1-pi2)/((1-pi1)*pi2) # find odds ratio for each generate
> quantile(or, c(0.025, 0.975))
    2.5%;     97.5%;
3.277981   1361813 # approx. posterior interval for odds ratio

> quantile(pi1 - pi2, c(0.025, 0.975))
  2.5%;    97.5%;
0.0990   0.9933 # approx. posterior interval for difference of proportions

> mean(pi1 < pi2)
[1] 0.0059  # approx. posterior P(pi1 < pi2), (control better than exper.)
-------------------------------------------------------------------------------
```

The posterior $P(\pi_1 \leq \pi_2) = 0.006$ with beta(0.5, 0.5) priors and 0.011 with uniform priors. The evidence is strong that the experimental treatment is better than the control.

2.7 ASSOCIATION IN THREE-WAY TABLES

An important part of most research studies is the choice of control variables. In studying the effect of an explanatory variable X on a response variable Y, we should adjust for confounding variables that can influence that relationship because they are associated both with X and with Y. Otherwise, an observed XY association may merely reflect associations of those variables with X and Y. This is especially vital for observational studies, for which one cannot remove effects of such variables by randomly assigning subjects to different treatments.

In a study of the effects on a nonsmoker of living with a smoker, a cross-sectional study might compare lung cancer rates between nonsmokers whose spouses smoke and nonsmokers whose spouses do not smoke. In doing so, the study should attempt to control for age, socioeconomic status, or other factors that might relate both to whether one's spouse smokes and to whether one has lung cancer. A statistical control would adjust for such variables while studying the association. Without such controls, results will have limited usefulness. Suppose that spouses of nonsmokers tend to be younger than spouses of smokers and that younger people are less likely to have lung cancer. Then, a lower proportion of lung cancer cases among nonsmoker spouses may merely reflect their lower average age and not an effect of passive smoking.

Including control variables in an analysis requires a multivariate rather than a bivariate analysis. We illustrate basic concepts for a single control variable Z, which is categorical. A three-way contingency table displays counts for the three variables.

2.7.1 Partial Tables

For three-way contingency tables, two-way *partial tables* cross-classify X and Y at separate categories for Z. They display the *XY* relationship while removing the effect of Z by holding its value constant.

The associations in partial tables are called *conditional associations*, because they refer to the effect of X on Y conditional on Z being constant. Conditional associations in partial tables can be quite different from associations in the *XY marginal table* that results from combining the partial tables. The next example illustrates.

2.7.2 Example: Death Penalty Verdicts and Race

Table 2.9 is a $2 \times 2 \times 2$ contingency table from an article that studied the effects of racial characteristics on whether subjects convicted of homicide receive the death penalty. The 674 subjects were the defendants in indictments involving cases with multiple murders, in Florida during a 12-year period. The variables are $Y =$ death penalty verdict, $X =$ race of defendant, and $Z =$ race of victims. We study the effect of a defendant's race on the death penalty verdict, treating victims' race as a control variable. Table 2.9 has a 2×2 partial table relating defendant's race and the death penalty verdict at each category of victims' race.

Table 2.9 Death penalty verdict by defendant's race and victims' race.

Victims' Race	Defendant's Race	Death Penalty		Percentage
		Yes	No	Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

Source: M.L. Radelet and G.L. Pierce, *Florida Law Rev.* **43**: 1–34 (1991).
Reprinted with permission of the *Florida Law Review*.

For each combination of defendant's race and victims' race, Table 2.9 lists and Figure 2.4 displays the percentage of defendants who received the death penalty. When the victims were white, the death penalty was imposed $22.9\% - 11.3\% = 11.6\%$ more often for black defendants than for white defendants. When the victim was black, the death penalty was imposed $2.8\% - 0.0\% = 2.8\%$ more often for black defendants than for white defendants. Thus, *controlling* for victims' race, the percentage of *yes* death penalty verdicts was higher for black defendants than for white defendants.

The bottom portion of Table 2.9 displays the marginal table for defendant's race and the death penalty verdict. We obtain it by summing the cell counts in Table 2.9 over the two

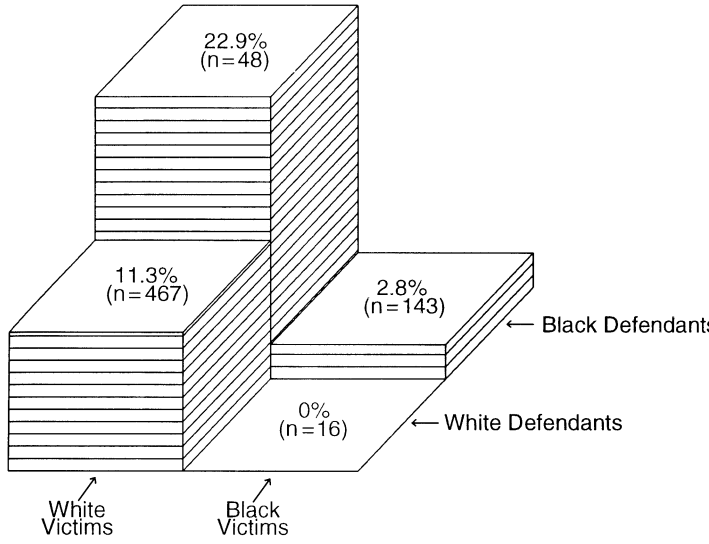


Figure 2.4 Proportion receiving the death penalty by defendant's race and victims' race.

categories of victims' race, thus combining the two partial tables (e.g., $11 + 4 = 15$). We see that, overall, 11.0% of white defendants and 7.9% of black defendants received the death penalty. *Ignoring* victims' race, the percentage of *yes* death penalty verdicts was lower for black defendants than for white defendants. The association reverses direction compared to the partial tables.

Why does the association between the death penalty verdict and defendant's race differ so much when we ignore rather than control victims' race? This relates to the nature of the association between the control variable, victims' race, and the other variables. First, the association between victims' race and defendant's race was extremely strong. The marginal table relating these variables has odds ratio 87.0. The odds that a white defendant had white victims were estimated to be 87.0 times the odds that a black defendant had white victims. Second, regardless of defendant's race, the death penalty was considerably more likely when the victims were white than when the victims were black. Therefore, whites were tending to kill whites and killing whites was more likely to result in the death penalty. This suggests that the marginal association should show a greater tendency for white defendants to receive the death penalty than do the conditional associations. In fact, Table 2.9 shows this pattern.

2.7.3 Simpson's Paradox

The result that a marginal association can have a different direction from the conditional associations is called *Simpson's paradox*. This result applies to quantitative as well as categorical variables. For example, quantitative variables X and Y can have a positive correlation, yet a negative partial correlation after adjusting for Z.

Figure 2.5 shows why Simpson's paradox happens. For each defendant's race, the figure plots the proportion receiving the death penalty at each victims' race. Each proportion is labeled by a letter symbol giving the victims' race category. Surrounding each observation is a circle having area proportional to the number of observations at that combination of defendant's race and victims' race. For instance, the W in the largest circle represents a

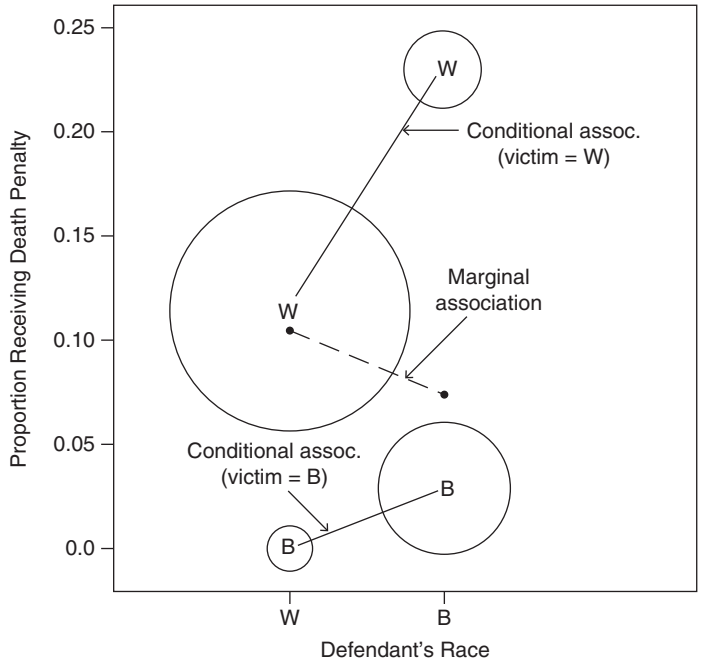


Figure 2.5 Proportion receiving the death penalty by defendant's race, controlling and ignoring victims' race.

proportion of 0.113 receiving the death penalty for cases with white defendants and white victims. That circle is largest because the number of cases at that combination ($53 + 414 = 467$) is larger than at the other three combinations. The next largest circle relates to cases in which blacks kill blacks.

To control for victims' race, we compare circles having the same victims' race letter at their centers. The line connecting the two W circles has a positive slope, as does the line connecting the two B circles. Controlling for victims' race, this reflects a higher chance of the death penalty for black defendants than white defendants. When we add results across victims' race to get a summary result for the marginal effect of defendant's race on the death penalty verdict, the larger circles having the greater number of cases have greater influence. Thus, the summary proportions for each defendant's race, marked on the figure by periods, fall closer to the center of the larger circles than the smaller circles. A line connecting the summary marginal proportions has negative slope. This indicates that, overall, white defendants are more likely than black defendants to receive the death penalty.

2.7.4 Conditional and Marginal Odds Ratios

Conditional associations, like marginal associations, can be described using odds ratios. We refer to odds ratios for partial tables as *conditional odds ratios*. For binary X and Y, within category k of Z, let $\theta_{XY(k)}$ denote the odds ratio between X and Y computed for the true probabilities.

From Table 2.9, the estimated conditional odds ratio between defendant's race and the death penalty in the first partial table, for which victims' race is white, equals $\hat{\theta}_{XY(1)} = (53 \times 37)/(414 \times 11) = 0.43$. The sample odds for white defendants receiving the death penalty were 43% of the sample odds for black defendants. In the second partial table, for which victim's race is black, $\hat{\theta}_{XY(2)} = (0 \times 139)/(16 \times 4) = 0.0$, because the death penalty was never given to white defendants having black victims.

The conditional odds ratios can be quite different from the marginal odds ratio. The marginal odds ratio for defendant's race and the death penalty uses the 2×2 marginal table in Table 2.9, ignoring instead of controlling victims' race. The estimate equals $(53 \times 176)/(430 \times 15) = 1.45$. The sample odds of the death penalty were 45% higher for white defendants than for black defendants. Yet, we just observed that those odds were smaller for a white defendant than for a black defendant, for each victims' race. This reversal in the association when we control for victims' race illustrates Simpson's paradox.

If the population has X and Y independent in each partial table, then X and Y are said to be *conditionally independent, given Z.* All conditional odds ratios between X and Y then equal 1. Conditional independence of X and Y, given Z, does not imply marginal independence of X and Y. That is, when odds ratios between X and Y equal 1 at each category of Z, the marginal odds ratio may differ from 1. An association can exist between two variables but completely disappear when we adjust for another variable. Exercise 2.28 shows an example.

2.7.5 Homogeneous Association

Two binary variables X and Y satisfy *homogeneous association* when

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots,$$

that is, when all $\{\theta_{XY(k)}\}$ are identical. Conditional independence of X and Y is the special case in which each conditional odds ratio equals 1.0.

For $X =$ smoking (yes, no), $Y =$ lung cancer (yes, no), and $Z =$ age (<45, 45-65, >65), suppose $\theta_{XY(1)} = 1.2$, $\theta_{XY(2)} = 4.8$, and $\theta_{XY(3)} = 12.2$. Then, smoking has a weak effect on lung cancer for young people, but the effect strengthens considerably with age. The *XY* association is not homogeneous, because the *XY* conditional odds ratio changes across levels of Z, but if Z is gender and $\theta_{XY(1)} = \theta_{XY(2)} = 4.0$, then homogeneous association occurs.

In a three-way table, homogeneous XY association means that any conditional odds ratio formed using two categories of X and two categories of Y is the same at each category of Z. Inference about associations in multi-way contingency tables is best handled in the context of models. Sections 4.3 and 7.1 introduce models that have the property of homogeneous association. We will see there and in Section 5.2 how to use the data to judge whether conditional independence or homogeneous association are plausible.

EXERCISES

2.1 The PSA blood test is designed to detect prostate cancer. Suppose that of men who have this disease, the test fails to detect prostate cancer in 1 in 4, and of men who do not have it, 1 in 10 have positive test results (so-called false-positive results).

Let C ($\bar{C}$) denote the event of having (not having) prostate cancer and let $+$ ($-$) denote a positive (negative) test result.

a. Which is true: $P(- \mid C) = 1/4$ or $P(C \mid -) = 1/4$? $P(\bar{C} \mid +) = 1/10$ or $P(+ \mid \bar{C}) = 1/10$?

b. Find the sensitivity and specificity of this test.

c. Of men who take the PSA test, suppose $P(C) = 0.04$. Find the cell probabilities in the 2×2 table for the joint distribution that cross-classifies $Y =$ diagnosis with $X =$ true disease status.

d. Using (**c**), find the marginal distribution for the diagnosis and show that $P(C \mid +) = 0.238$. (In fact, the National Cancer Institute estimates that only about 25% of men who have a slightly elevated PSA level, 4–10 ng/mL, actually have prostate cancer.[18])

2.2 For diagnostic testing, let $X =$ true status ($1 =$ disease, $2 =$ no disease) and $Y =$ diagnosis ($1 =$ positive, $2 =$ negative). Let $\pi_1 = P(Y = 1 \mid X = 1)$ and $\pi_2 = P(Y = 1 \mid X = 2)$. Let γ denote the probability that a subject has the disease.

a. Given that the diagnosis is positive, use Bayes' Theorem to show that the probability a subject truly has the disease is

$$P(X = 1 \mid Y = 1) = \pi_1\gamma/[\pi_1\gamma + \pi_2(1 - \gamma)].$$

b. For mammograms for detecting breast cancer, suppose $\gamma = 0.01$, sensitivity $= \pi_1 = 0.86$, and specificity $1 - \pi_2 = 0.88$. Find the positive predictive value.

c. To better understand the answer in (**b**), find the joint probabilities for the 2×2 cross-classification of X and Y. Discuss their relative sizes in the two cells that refer to a positive test result.

2.3 According to recent UN figures, the annual gun homicide rate is 62.4 per one million residents in the US and 1.3 per one million residents in Britain. Compare the proportion of residents killed annually by guns using the (**a**) difference of proportions, (**b**) relative risk. Which measure is more useful for describing the strength of association? Why?

2.4 The opening 2018 World Cup odds against being the winning team specified by `espn.com` were $9/2$ for Germany, $5/1$ for Brazil, $11/2$ for France, $20/1$ for England, and $7/1$ for Spain. Find the corresponding prior probabilities of winning for these five teams.

2.5 Consider the following two studies reported in the *New York Times*:

a. A British study reported that of smokers who get lung cancer, "women were 1.7 times more vulnerable than men to get small-cell lung cancer." Is 1.7 an odds ratio or a relative risk?

b. A National Cancer Institute study about tamoxifen and breast cancer reported that the women taking the drug were 45% less likely to experience invasive breast cancer compared to the women taking placebo. Find the relative risk for (i) those taking the drug compared to those taking placebo, (ii) those taking placebo compared to those taking the drug.

[18] See `www.cancer.gov/types/prostate/psa-fact-sheet`.

2.6 Finding and interpreting measures comparing proportions:

a. An observational study[19] of patients hospitalized for gunshot wounds in the US between 2004 and 2013 classified the intent of the gunshot using categories (assault, unintentional, suicide, undetermined, law enforcement, terrorism), with counts (179,793, 65,502, 24,624, 17,401, 5266, 10). Compare proportions for the categories *unintentional* and *terrorism* by a difference and by a ratio. Interpret.

b. According to a 2015 study by the Pew Research Center (`www.people-press.org`), the percentage of Americans who favor allowing gays and lesbians to marry legally was 81% for Democrats who identified themselves as liberal and 22% for Republicans who identified themselves as conservative. Identify the two variables and find the odds ratio between them.

2.7 For adults who sailed on the Titanic on its fateful voyage, the odds ratio[20] between gender (female, male) and survival (yes, no) was 11.4.

a. What is wrong with the interpretation, "The probability of survival for females was 11.4 times that for males?" Give the correct interpretation.

b. The odds of survival for females equaled 2.9. For each gender, find the proportion who survived. Find the value of RR in the interpretation, "The probability of survival for females was RR times that for males."

2.8 A research study estimated that under a certain condition, the probability a subject would be referred for heart catheterization was 0.906 for whites and 0.847 for blacks.

a. A press release about the study stated that the odds of referral for cardiac catheterization for blacks are 60% of the odds for whites. Explain how they obtained 60% (more accurately, 57%).

b. An Associated Press story[21] that described the study stated "Doctors were only 60% as likely to order cardiac catheterization for blacks as for whites." What is wrong with this interpretation? Give the correct percentage for this interpretation. (In stating results to the general public, it is better to use the relative risk than the odds ratio. It is simpler to understand and less likely to be misinterpreted.)

2.9 A 20-year study of British male physicians[22] noted that the proportion who died from lung cancer was 0.00140 per year for cigarette smokers and 0.00010 per year for nonsmokers. The proportion who died from heart disease was 0.00669 for smokers and 0.00413 for nonsmokers. Describe the association of smoking with lung cancer and with heart disease, using the difference of proportions and the odds ratio. Interpret. Which response (lung cancer or heart disease) is more strongly related to cigarette smoking, in terms of the reduction in deaths that could occur with an absence of smoking?

2.10 Table 2.10 shows fatality results for drivers and passengers in auto accidents in Florida in 2015, according to whether the person was wearing a shoulder and lap belt restraint versus not using one. Find and interpret the odds ratio.

2.11 Table 2.11 cross-classifies votes in the 2008 and 2012 US Presidential elections. Estimate and find a 95% confidence interval for the population odds ratio. Interpret.

[19] By A. Cook, T. Osler, D. Hosmer, et al., *Injury* **48**: 621–627 (2017).
[20] For data, see R. Dawson, *J. Statist. Educ.* **3** (1995).
[21] For details, see *N. Engl. J. Medic.* **341**: 279–283 (1999).
[22] By R. Doll and R. Peto, *British Med. J.* **2**: 1525–1536 (1976).

Table 2.10 Data for exercise 2.10 on auto accidents.

	Injury		
Restraint Use	Fatal	Nonfatal	Total
No	433	8049	8482
Yes	570	554,883	555,453

Source: Florida Department of Highway Safety and Motor Vehicles.

Table 2.11 Data on presidential votes, for exercise 2.11.

	Vote in 2012	
Vote in 2008	Obama	Romney
Obama	802	53
McCain	34	494

Source: 2014 General Social Survey.

2.12 Data posted at the FBI website (`www.fbi.gov`) indicated that of all blacks slain in 2015, 92% were slain by blacks, and of all whites slain in 2015, 93% were slain by whites. Let Y denote race of victim and X denote race of murderer. Which conditional distribution do these statistics refer to, Y given X or X given Y? Find and interpret the odds ratio.

2.13 Refer to Table 2.1 about belief in an afterlife. Conduct a test of statistical independence. Report the P-value and interpret.

2.14 A poll by Louis Harris and Associates of 1249 adult Americans indicated that 36% believe in ghosts and 37% believe in astrology. Can you compare the proportions using inferential methods for independent binomial samples? If yes, do so. If not, explain why not.

2.15 An article[23] summarized results from the Nurses' Health Study and the Health Professionals Follow-Up Study. The article reported (with RR = relative risk) that "Compared with nonregular use, regular aspirin use was associated with lower risk of overall cancer (RR 0.97; 95% CI 0.94, 0.99), which was primarily due to a lower incidence of gastrointestinal cancers, especially colorectal cancers (RR 0.81; 95% CI 0.75, 0.88)."

a. Identify the response variables and the explanatory variable for these two results. Explain how to interpret the confidence interval about colorectal cancers.

b. Would the association with overall cancer be considered (i) significant or nonsignificant? (ii) strong or weak? Explain.

2.16 Table 2.12 shows data from a General Social Survey cross-classifying a person's perceived happiness with their family income. The table displays the observed and expected cell counts and the standardized residuals for testing independence.

a. For testing independence, $X^2 = 73.4$. Report the *df* value and the P-value, and interpret.

b. Interpret the standardized residuals in the corner cells having (i) counts 21 and 83, (ii) counts 110 and 94.

[23] By Y. Cao et al., *JAMA Oncology* **2**: 762–769 (2016).

Table 2.12 Data for Exercise 2.16, with estimated expected frequencies and standardized residuals.

	Happiness		
Income	Not Too Happy	Pretty Happy	Very Happy
Above average	21 (35.8) −2.973	159 (166.1) −0.947	110 (88.1) 3.144
Average	53 (79.7) −4.403	372 (370.0) 0.224	221 (196.4) 2.907
Below average	94 (52.5) 7.368	249 (244.0) 0.595	83 (129.5) −5.907

2.17 Table 2.13 is based on data from the 2016 General Social Survey.

a. Test the null hypothesis of independence between political party identification and race. Interpret.

b. Use standardized residuals to describe the evidence.

c. Partition chi-squared into two components, the first of which compares the races on the (Democrat, Republican) choice. Interpret the quite different results for the two cases.

Table 2.13 Data for Exercise 2.17.

	Political Party Identification		
Race	Democrat	Republican	Independent
White	871	821	336
Black	347	42	83

2.18 Each subject in a sample of 100 men and 100 women is asked to indicate which of the following factors (one or more) are responsible for increases in teenage crime: A – the increasing gap in income between the rich and poor, B – the increase in the percentage of single-parent families, C – insufficient time that parents spend with their children. A cross-classification of the responses by gender is

```
-----------------------
Gender    A    B    C
Men       60   81   75
Women     75   87   86
-----------------------
```

Is it valid to apply the chi-squared test of independence to this table? Why or why not? Explain how this table actually provides information needed to cross-classify gender with each of three variables. Construct the contingency table relating gender to opinion about whether factor A is responsible for increases in teenage crime.

2.19 Table 2.14 is from a recent General Social Survey. For these data, $X^2 = 69.2$. Write a short report summarizing inference. In your report, mention an alternative test of independence that is relevant for these data.

Table 2.14 Table for Exercise 2.19, with standardized residuals.

	Religious Beliefs		
Highest Degree	Fundamentalist	Moderate	Liberal
Less that High School	178 (4.5)	138 (−2.6)	108 (−1.9)
High School or Junior College	570 (2.6)	648 (1.3)	442 (−4.0)
Bachelor or Graduate	138 (−6.8)	252 (0.7)	252 (6.3)

2.20 Formula (2.3) has alternative formula $X^2 = n\sum(\hat{\pi}_{ij} - \hat{\pi}_{i+}\hat{\pi}_{+j})^2/\hat{\pi}_{i+}\hat{\pi}_{+j}$. Explain why, for particular $\{\hat{\pi}_{ij}\}$, X^2 is large when n is sufficiently large, regardless of whether the association is practically important. Hence, chi-squared tests merely indicate the degree of evidence against a hypothesis and do not describe the strength of association.

2.21 A GSS that cross-classified income in thousands of dollars (<5, 5–15, 15–25, >25) by job satisfaction (very dissatisfied, a little satisfied, moderately satisfied, very satisfied) for black Americans produced a 4×4 table having counts, by row, (2, 4, 13, 3 / 2, 6, 22, 4 / 0, 1, 15, 8 / 0, 3, 13, 8).

a. Test independence of job satisfaction and income using X^2. Interpret and explain the deficiency of this test for these data. Find the standardized residuals. Do they suggest any association pattern?

b. Conduct a test that treats the variables in a quantitative manner, using scores (3, 10, 20, 35) for income and (1, 3, 4, 5) for job satisfaction. Explain why results differ so much from part (**a**).

2.22 A study (B. Kristensen et al., *J. Intern. Med.* **232**: 237–245 (1992)) considered the effect of prednisolone on severe hypercalcaemia in women with metastatic breast cancer. Of 30 patients, 15 were randomly selected to receive prednisolone and the other 15 formed a control group. Normalization in their level of serum-ionized calcium was achieved by 7 of the 15 prednisolone-treated patients and by 0 of the 15 patients in the control group. Use Fisher's exact test to find a P-value for testing whether results were significantly better for treatment than control. Interpret.

2.23 Table 2.15 contains results of a study comparing radiation therapy with surgery in treating cancer of the larynx. Some R output follows:

```
------------------------------------------------------------------------------
> fisher.test(matrix(c(21,2,15,3),ncol=2,byrow=TRUE),alternative="two.sided")
p-value = 0.6384
> fisher.test(matrix(c(21,2,15,3),ncol=2,byrow=TRUE),alternative="greater")
p-value = 0.3808
> fisher.test(matrix(c(21,2,15,3),ncol=2,byrow=TRUE),alternative="less")
p-value = 0.8947
------------------------------------------------------------------------------
```

Table 2.15 Data for Exercise 2.23.

	Cancer Controlled	Cancer Not Controlled
Surgery	21	2
Radiation therapy	15	3

Source: W. Mendenhall et al., *Int. J. Radiat. Oncol. Biol. Phys.* **10**: 357–363 (1984), with permission from Elsevier Science Ltd.

a. Report and interpret the P-value for Fisher's exact test with (**i**) H_a: $\theta > 1$, and (**ii**) H_a: $\theta \neq 1$.

b. Obtain and interpret the mid P-value for H_a: $\theta \neq 1$ and find the corresponding confidence interval based on mid P-values. Give advantages of this type of P-value, compared to the ordinary one.

2.24 a. At each age level, the death rate is higher in South Carolina than in Maine, but overall the death rate is higher in Maine.[24] Explain how this could be possible.

b. Smith and Jones are baseball players. Smith had a higher batting average than Jones in 2005 and 2006. Is it possible that for the combined data for these two years, Jones had the higher batting average? Explain, and illustrate using data.

2.25 In murder trials in 20 Florida counties during 1976 and 1977, the death penalty was given in 19 out of 151 cases in which a white killed a white, in 0 out of 9 cases in which a white killed a black, in 11 out of 63 cases in which a black killed a white, and in 6 out of 103 cases in which a black killed a black.[25]

a. Exhibit the data as a three-way contingency table. Construct the partial tables needed to study the conditional association between defendant's race and the death penalty verdict. Find and interpret the sample conditional odds ratios.

b. Find and interpret the sample marginal odds ratio between defendant's race and the death penalty verdict. Do these data exhibit Simpson's paradox? Explain its cause.

2.26 Give an example of three variables X, Y, and Z, for which you expect X and Y to be marginally associated but conditionally independent, controlling for Z.

2.27 Based on murder rates in the United States, the Associated Press reported that the probability a newborn child has of eventually being a murder victim is 0.0263 for nonwhite males, 0.0049 for white males, 0.0072 for nonwhite females, and 0.0023 for white females. Find the conditional odds ratios between race and whether a murder victim, given gender. Interpret.

2.28 The expected frequencies in Table 2.16 show a hypothetical relationship among three variables: Y = response, X = drug treatment, and Z = clinic. Show that X and

Table 2.16 Expected frequencies illustrating that conditional independence does not imply marginal independence.

Clinic	Drug Treatment	Response	
		Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32

[24] For data, see H. Wainer, *Chance* **12**: 44 (1999).
[25] M. Radelet, *Amer. Sociol. Rev.* **46**: 918–927 (1981).

Y are conditionally independent, given Z, but marginally associated. Explain how the marginal XY association can be so different from its conditional association, using the values of the conditional *XZ* and *YZ* odds ratios. Explain why it would be misleading to study only the marginal table and conclude that successes are more likely with treatment A than with treatment B.

2.29 Refer to Table 2.1 about belief in an afterlife.

a. Treating the data as independent binomial samples and using beta(0.5, 0.5) prior distributions, find the posterior mean estimates for the probabilities of believing in an afterlife.

b. Find and interpret (i) 95% posterior intervals for the difference of proportions and the odds ratio; (ii) the posterior probability that belief in an afterlife is more probable for women than for men.

2.30 True or false?

a. A 95% confidence interval for the odds ratio between MI (yes, no) and treatment (placebo, aspirin) is (1.44, 2.33). If we form the table with aspirin in the first row (instead of placebo), the confidence interval is $(1/2.33, 1/1.44) = (0.43, 0.69)$.

b. A survey of college students analyzes the association between opinion about whether it should be legal to (1) use marijuana, (2) drink alcohol if you are 18 years old. We may get a different odds ratio value if we treat marijuana use as the response variable than if we treat alcohol use as the response variable.

c. Interchanging two rows or interchanging two columns in a contingency table has no effect on the value of the X^2 or G^2 chi-squared statistics. Thus, these tests treat both the rows and the columns of the contingency table as nominal scale, and if either or both variables are ordinal, the test ignores that information.

d. Suppose that income (high, low) and gender are conditionally independent, given the type of job (secretarial, construction, service, professional, ...). Then, income and gender are also independent in the 2×2 marginal table.

e. According to the Pew Research Center (`www.people-press.org`), when adults in the US were asked in 2010 whether there is solid evidence that the average temperature on Earth has been getting warmer over the past few decades, the estimated odds of a *yes* response for a Democrat was 2.96 times higher than for an Independent, and it was 2.08 times higher for an Independent than for a Republican. The estimated odds ratio between opinion on global warming and whether one is a Democrat or a Republican equals $2.96 \times 2.08 = 6.2$.

CHAPTER 3

GENERALIZED LINEAR MODELS

The methods presented in Chapter 2 for analyzing contingency tables help us investigate effects of a categorical explanatory variable on a categorical response variable. The rest of this book uses *models* as the basis of such analyses. In fact, the methods of Chapter 2 also result from analyzing effects in certain models. However, models can handle more complicated situations, such as analyzing simultaneously the effects of several explanatory variables, which can be categorical or quantitative or both.

A good-fitting model has several benefits. The structural form of the model describes the patterns of association and interaction. The sizes of the model parameters determine the strength and importance of the effects. Inferences about the parameters evaluate which explanatory variables truly are associated with the response variable Y, while adjusting for effects of other variables, such as possible confounding variables. Finally, the model's predicted values smooth the data and provide improved estimates of the mean of Y at possible explanatory variable values.

The models that this book presents are *generalized linear models*. The acronym *GLM* is shorthand for *g*eneralized *l*inear *m*odel. This broad class of models includes ordinary regression and analysis of variance (ANOVA) models for continuous response variables as well as models for categorical response variables. After presenting the three components of GLMs, this chapter introduces models for categorical and other discrete response variables. GLMs for categorical responses include *logistic regression models*, which are the main focus of Chapters 4 to 6. GLMs for discrete response variables for which the outcome is a count

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

include *loglinear models*, the subject of Chapter 7. This chapter also introduces inference methods, model checking, and model fitting for GLMs.

3.1 COMPONENTS OF A GENERALIZED LINEAR MODEL

Generalized linear models have three components: The *random component* identifies the response variable Y and its probability distribution. The *linear predictor* specifies the explanatory variables through a prediction equation that has linear form. The *link function* specifies a function of $E(Y)$ that the GLM relates to the linear predictor.

3.1.1 Random Component

The *random component* of a GLM identifies the response variable Y and assumes a probability distribution for it. Standard GLMs treat the n observations on Y as independent. We denote those observations by $(y_1, \ldots, y_n)$.

In many applications, the observations are binary, such as *success* or *failure*. More generally, each y_i might be the number of successes out of a certain fixed number of trials. In either case, we assume a *binomial* distribution for Y. In some applications, each observation is a count. We then assume a distribution for Y that is defined on all the nonnegative integers, usually the *Poisson* or the *negative binomial*. If each observation is continuous, such as a subject's weight in a dietary study, we might assume a *normal* or a *gamma* distribution for Y.

3.1.2 Linear Predictor

The *linear predictor* of a GLM specifies the explanatory variables. The name reflects that the variables enter linearly as predictors on the right-hand side of the model equation, in the form

$$\alpha + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Statistical inference for the model conditions on the observed values of the explanatory variables, treating them as fixed rather than as random variables.

As in ordinary regression, some $\{x_j\}$ can be based on others in the model. For example, perhaps $x_3 = x_1 x_2$, to allow interaction between x_1 and x_2 in their effects on Y.

3.1.3 Link Function

The expected value $\mu = E(Y)$ of the probability distribution of Y has a value that varies according to values of the explanatory variables. The third component of a GLM, the *link function*, specifies a function g that relates μ to the linear predictor as

$$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p.$$

The link function g connects the random component with the linear predictor function of the explanatory variables.

The simplest link function is $g(\mu) = \mu$. This models the mean directly and is called the *identity link function*. It specifies a linear model for the mean response,

$$\mu = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p.$$

This is the form of ordinary linear models for quantitative response variables.

Other link functions permit μ to be nonlinearly related to the explanatory variables. For instance, the link function $g(\mu) = \log(\mu)$ models the log of the mean. The log function applies to positive numbers, so the *log link function* is appropriate when μ cannot be negative, such as with count data. A GLM that uses the log link is called a *loglinear model*. It has the form

$$\log(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p.$$

The link function $g(\mu) = \log[\mu/(1-\mu)]$ models the log of an odds. It is appropriate when μ is between 0 and 1, such as a probability. This is called the *logit link function*. A GLM that uses the logit link function is called a *logistic regression model*.

Each potential probability distribution for Y has a special function of the mean that is its *natural parameter*. For the normal distribution, it is the mean itself. For the binomial, the natural parameter is the logit of the success probability. For the Poisson, it is the log of the mean. The link function that uses the natural parameter as $g(\mu)$ in the GLM is called the *canonical link function*. Although other link functions are possible, by default, software for GLMs uses the canonical link functions.

3.1.4 Ordinary Linear Model: GLM with Normal Random Component

Ordinary linear models, such as are the basis of regression analysis, are special cases of GLMs. They assume a normal distribution for Y and model its mean directly, using the identity link function, $g(\mu) = \mu$. A GLM generalizes ordinary linear models in two ways: First, it allows Y to have a distribution other than the normal. Second, it allows modeling some function of the mean. Both generalizations are important for categorical data.

Historically, early analyses of discrete response variables often attempted to transform Y so that it has an approximately normal distribution, with constant variance. Then, ordinary linear modeling methods using least squares are applicable. In practice, this is often not possible. With the theory and methodology of GLMs, it is unnecessary to transform data so that methods for normal responses apply. The GLM fitting process uses ML methods for our choice of a random component, and we are not restricted to normality for that choice. The GLM choice of link function is separate from the choice of random component. It is chosen to yield a linear predictor form for the effects, not to transform the data to produce normality or to stabilize the variance.

GLMs unify a wide variety of statistical methods. Regression models and models for categorical data are special cases of one supermodel. In fact, the same algorithm yields ML estimates of parameters for all GLMs. This algorithm, presented in Section 3.5, is the basis of software for fitting GLMs, such as the `glm` function in `R`, the `glm` command in Stata, and PROC GENMOD in SAS.

The next two sections illustrate the GLM components by introducing two important GLMs for discrete response variables: logistic regression models for binary data and loglinear models for count data.

3.2 GENERALIZED LINEAR MODELS FOR BINARY DATA

Many categorical response variables have only two categories: for example, (yes, no) categories for whether you take public transportation today, whether you are employed, and your opinion about whether global warming is truly occurring. We denote the two possible outcomes for a binary response variable Y by 1 (*success*) and 0 (*failure*). The distribution of Y is specified by probabilities $P(Y = 1) = \pi$ of success and $P(Y = 0) = (1 - \pi)$ of failure. Its mean is $E(Y) = \pi$. For n independent observations, the number of successes has the binomial distribution formula (1.1) (in Section 1.2.1) specified by the index n and parameter π. Each binary observation is a binomial variate with $n = 1$.

3.2.1 Linear Probability Model

When Y is binary, we could mimic ordinary linear modeling and use an identity link function,

$$P(Y = 1) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p.$$

This is called a *linear probability model*, because the probability of success changes linearly in each explanatory variable. For example, the parameter β_1 represents the change in $P(Y = 1)$ per unit change in x_1, adjusting for the other variables. This model is a GLM with a binomial random component and identity link function.

This model is simple and easy to interpret. However, it has a structural defect. Probabilities fall between 0 and 1, whereas linear predictors take values over the entire real line. Because of this, this model can predict values $P(Y = 1) < 0$ and $P(Y = 1) > 1$ for some values of the explanatory variables. It can fit adequately over a restricted range of explanatory variable values and more generally can provide useful summaries of effects. For most applications, however, especially with multiple explanatory variables, we need a more complex model form to be able to estimate $P(Y = 1)$ well.

3.2.2 Logistic Regression Model

Effects of explanatory variables on $P(Y = 1)$ are usually nonlinear rather than linear. A fixed change in an x may have less impact when $P(Y = 1)$ is near 0 or 1 than when $P(Y = 1)$ is near the middle of its range. In the purchase of an automobile, for instance, consider modeling the choice between buying *new* (1) or *used* (0) in terms of $x =$ annual family income. An increase of \$10,000 in x would likely have less effect when $x =$ \$1,000,000, for which $P(Y = 1)$ is near 1, than when $x =$ \$50,000.

In practice, $P(Y = 1)$ often either increases continuously or decreases continuously as an x increases. The S-shaped curves displayed in Figure 3.1 might then portray shapes for the relationship with a single explanatory variable. An important mathematical function with this shape has the formula

$$P(Y = 1) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}, \tag{3.1}$$

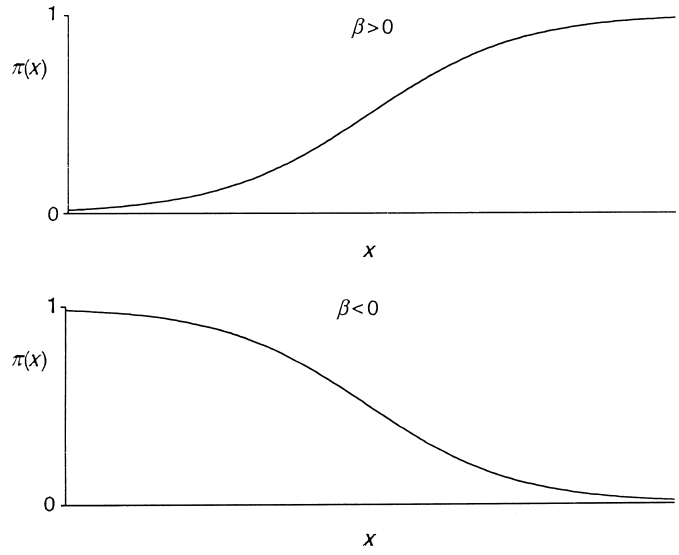


Figure 3.1 Logistic regression functions.

using the exponential function. This is called the *logistic regression* function. We will see in Chapter 4 that the corresponding logistic regression model form with multiple explanatory variables is

$$\log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p.$$

The logistic regression model is a special case of a GLM. The random component for the (success, failure) outcomes has a binomial distribution. The link function of $\pi = P(Y = 1)$ is the *logit* function, $\log[\pi/(1-\pi)]$, symbolized by "logit(π)." Logistic regression models are often called *logit models*. Whereas $P(Y = 1)$ is restricted to the 0 to 1 range, the logit can be any real number. The real numbers are also the potential range for linear predictors. This model therefore does not have the structural limitation that the linear probability model has.

The parameter β in (3.1) determines the rate of increase or decrease of the curve. When $\beta > 0$, $P(Y = 1)$ increases as x increases, as in Figure 3.1a. When $\beta < 0$, $P(Y = 1)$ decreases as x increases, as in Figure 3.1b. The magnitude of β determines how fast the curve increases or decreases. As $|\beta|$ increases, the curve has a steeper rate of change.

3.2.3 Example: Snoring and Heart Disease

Table 3.1 is based on an epidemiological survey to investigate snoring as a possible risk factor for heart disease. The subjects were classified according to their snoring level, as reported by their spouses. We treat the rows of the table as four independent binomial samples. We use scores (0, 2, 4, 5) for $x =$ snoring level, treating the last two snoring categories as closer than the other adjacent pairs.

Table 3.1 Relationship between snoring and heart disease, with model fits for the proportion of *yes* responses.

Snoring	Heart Disease		Proportion	Linear	Logistic
	Yes	No	Yes	Fit	Fit
Never	24	1355	0.017	0.017	0.021
Occasional	35	603	0.055	0.057	0.044
Nearly every night	21	192	0.099	0.096	0.093
Every night	30	224	0.118	0.116	0.132

Source of data: P.G. Norton and E.V. Dunn, *Brit. Med. J.* **291**: 630–632 (1985), published by BMJ Publishing Group.

The linear probability model states that the probability of heart disease is a linear function of the snoring level x. Software for GLMs reports the ML model fit,

$$\hat{P}(Y = 1) = 0.017 + 0.020x.$$

The model interpretation is simple. The estimated probability of heart disease is 0.017 for nonsnorers ($x = 0$), it increases 2(0.020) $= 0.04$ for occasional snorers ($x = 2$), another 0.04 for those who snore nearly every night ($x = 4$), and another 0.02 for those who always snore ($x = 5$).

The estimated values of $E(Y)$ for a GLM are called *fitted values*. Software for GLMs can show them. Table 3.1 shows the sample proportions and the fitted values for the linear probability model. Figure 3.2 plots the sample proportions and the fitted values. The table and graph suggest that the model fits these data well.

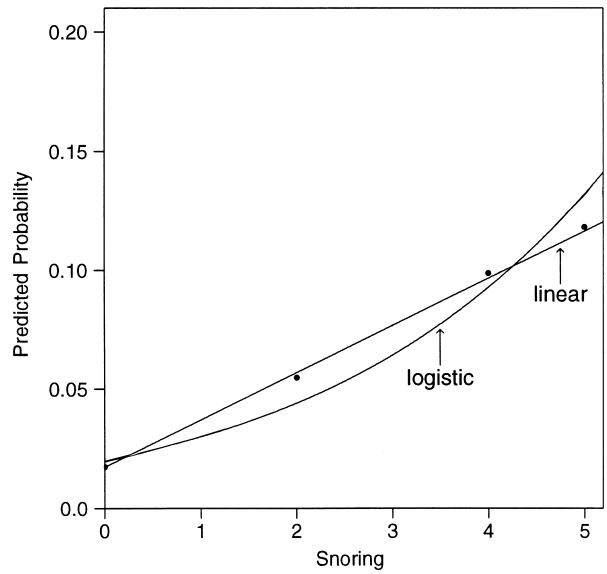


Figure 3.2 Fit of models for snoring and heart disease data.

For the logistic regression model, GLM software reports the ML fit,

$$\text{logit}[\hat{P}(Y = 1)] = -3.866 + 0.397x.$$

Since $\hat{\beta} = 0.397 > 0$, the estimated probability of heart disease increases as the snoring level increases. Chapter 4 presents ways of interpreting the model. Table 3.1 also reports its fitted values and Figure 3.2 displays the fit. The fit is close to linear over this rather narrow range of estimated probabilities. Results are similar to those for the linear probability model.

3.2.4 Using R to Fit Generalized Linear Models for Binary Data

You can easily use R to fit the logistic regression model to the contingency table data of Table 3.1. The glm function can fit a wide variety of generalized linear models. In the following code, we created an explanatory variable x that takes values (0, 2, 4, 5) for the four rows of the table.

```
------------------------------------------------------------------------------
> Heart <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Heart.dat",
+                     header=TRUE)
> Heart # Heart data file at text website, in contingency table form
             snoring yes   no
1              never  24 1355
2         occasional  35  603
3 nearly_every_night  21  192
4        every_night  30  224
> # the following code fits logistic regression model to the data file
> library(dplyr) # to recode explanatory variable
> Heart$x <- recode(Heart$snoring, never = 0, occasional = 2,
+                   nearly_every_night = 4, every_night = 5)
> n <- Heart$yes + Heart$no # binomial sample sizes are the row totals
> fit <- glm(yes/n ~ x, family=binomial(link=logit), weights=n, data=Heart)
> # canonical link for binomial is logit, so "(link=logit)" not necessary
> # "weights" indicates sample proportion yes/n is based on n observations
> summary(fit)
             Estimate   Std. Error
(Intercept)  -3.86625      0.16621
x             0.39734      0.05001 # logistic ML estimate of beta is 0.397
> fitted(fit) # fitted values (probability estimates) at 4 levels of snoring
      1        2        3        4
0.02051  0.04430  0.09305  0.13244
------------------------------------------------------------------------------
```

The identity link function is not available in R for its glm function with the binomial random component but you can fit it using the following code, which yields the result $\hat{P}(Y = 1) = 0.017 + 0.020x$:

```
------------------------------------------------------------------------------
> fit2 <- glm(yes/n ~ x, family=quasi(link=identity, variance="mu(1-mu)"),
+             weights=n, data=Heart)
> summary(fit2, dispersion=1)
             Estimate  Std. Error
(Intercept)  0.017247    0.003451
x            0.019778    0.002805
------------------------------------------------------------------------------
```

3.2.5 Data Files: Ungrouped or Grouped Binary Data

A standard data file has a row for each subject (e.g., person) and a column for each variable. For a binary response variable, the data file has a column of 1 (success) and 0 (failure) values. An example is the `Clinical` data file shown in Section 1.6.2. We refer to such data as *ungrouped*. When the explanatory variables are all discrete, data files can alternatively show the observations as binomial counts at the various combinations of values of the explanatory variables. We will refer to such data as *grouped*.

Table 3.1 on snoring and heart disease showed results for $n = 2484$ people. For the 1379 subjects who reported never snoring, 24 had heart disease and 1355 did not. For a grouped data file, such as in the `R` analysis shown at the end of Section 3.2.4, a line in the data file reports these data as 24 cases of heart disease and 1355 not having heart disease. The full data file then has only 4 lines. The ungrouped data file has a separate line for each person. Therefore, for those who reported never snoring, 24 lines contain a 1 for heart disease and 1355 lines contain a 0 for heart disease. The data file then has 2484 lines, for a binary response y that is 1 or 0 according to the presence of heart disease. It looks like:

```
-------------------------------------------------------------------------
> Heart2 # use Head(Heart2) to display first 6 lines of data file
subject       snoring    y
      1         never    1 # 3 of the 2484 lines of Heart2 data file
      2         never    1 # at text website
...
   2484         every    0
-------------------------------------------------------------------------
```

When at least one explanatory variable is continuous, binary data are naturally ungrouped. An example is the `Crabs` data file analyzed below in Section 3.3.3, which has 173 lines of data for the 173 crabs, at various width and weight values.

Whether the data are in an ungrouped or grouped data file has no effect on the fit of a GLM for binary data. The ML estimates and SE values are the same for either type of data file.

3.3 GENERALIZED LINEAR MODELS FOR COUNTS AND RATES

Many discrete response variables have *counts* as possible outcomes. Examples are the number of devices you own that can access the internet, the number of sex partners you have had in your lifetime, and the number of imperfections on a silicon wafer used in manufacturing computer chips. Counts also occur in summarizing categorical variables with contingency tables.

3.3.1 Poisson Distribution for Counts

Some GLMs for count response variables assume a *Poisson distribution* for the random component. Like counts, Poisson random variables can take any nonnegative integer value. The Poisson probabilities are

$$P(y) = \frac{e^{-\mu}\mu^y}{y!} \quad \text{for} \quad y = 0, 1, 2, \ldots$$

We won't need to use this formula, but we will use some of its properties.

The Poisson distribution is unimodal and skewed to the right. It has a single parameter $\mu > 0$, which is both its mean and its variance. That is,

$$E(Y) = \text{var}(Y) = \mu, \qquad \sigma(Y) = \sqrt{\mu}.$$

Therefore, when the counts are larger, on the average, they also tend to be more variable. If $Y =$ number of concerts attended in the past year has a Poisson distribution, then we observe greater variability in y from person to person when $\mu = 10.4$ than when $\mu = 1.2$. The mode is the integer part of the mean (e.g., mode $= 10$ when $\mu = 10.4$). As μ increases, the skew decreases and the distribution becomes more bell-shaped.[1]

3.3.2 Poisson Loglinear Model

Poisson GLMs can use the identity link, but it is more common to model the log of the mean. Like the linear predictor, the log of the mean can take any real-number value. A *Poisson loglinear model* is a GLM that assumes a Poisson distribution for Y and uses the log link function.

For a single explanatory variable x, the Poisson loglinear model has form

$$\log \mu = \alpha + \beta x.$$

The mean satisfies the exponential relationship

$$\mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}. \tag{3.2}$$

A one-unit increase in x has a multiplicative impact of e^{β} on μ: the mean of Y at $x + 1$ equals the mean of Y at x multiplied by e^{β}. If $\beta = 0$, then $e^{\beta} = e^{0} = 1$ and the multiplicative factor is 1. Then, the mean of Y does not change as x changes. If $\beta > 0$, then $e^{\beta} > 1$, and the mean of Y increases as x increases. If $\beta < 0$, the mean decreases as x increases.

3.3.3 Example: Female Horseshoe Crabs and their Satellites

Table 3.2 comes from a study of nesting horseshoe crabs.[2] Each female horseshoe crab in the study had a male crab attached to her in her nest. The study investigated factors that affect whether the female crab had any other males, called *satellites*, residing near her.[3]

Table 3.2 Number of crab satellites by female's color, spine condition, shell width, and weight[a].

C	S	Wi	Wt	Sa	C	S	Wi	Wt	Sa	C	S	Wi	Wt	Sa
2	3	28.3	3.05	8	3	3	22.5	1.55	0	1	1	26.0	2.30	9
3	3	26.0	2.60	4	2	3	23.8	2.10	0	3	2	24.7	1.90	0
3	3	25.6	2.15	0	3	3	24.3	2.15	0	2	3	25.8	2.65	0
4	2	21.0	1.85	0	2	1	26.0	2.30	14	1	1	27.1	2.95	8

[a] C, color (1, medium light; 2, medium; 3, medium dark; 4, dark); S, spine condition (1, both good; 2, one broken; 3, both broken); Wi, shell width (cm); Wt, weight (kg); Sa, number of satellites. *Source:* Data courtesy of Jane Brockmann, University of Florida; study described in *Ethology* **102**: 1–21 (1996). The complete `Crabs` data file is at the text website.

[1] For graphs for various μ, see `https://en.wikipedia.org/wiki/Poisson_distribution`.

[2] See `https://en.wikipedia.org/wiki/Horseshoe_crab` for description and pictures.

[3] See `www.bluffton.com/wp-content/uploads/Spawning-horseshoe-crabs.jpg`.

The response outcome for each female crab is her number of satellites. An explanatory variable thought possibly to affect this was the female crab's carapace (shell) width, which is a summary of her size. In the sample, width had a mean of 26.3 cm and a standard deviation of 2.1 cm. Table 3.2 shows data, with four explanatory variables including width, for 12 of the 173 female horseshoe crabs.

Figure 3.3 plots the response counts against width. The substantial variability makes it difficult to discern a clear trend. Software has ways of smoothing the data, revealing any trends. *Generalized additive models*, introduced later in the text (Section 11.4) do this by providing more general structural form than GLMs. They find possibly complex functions of the explanatory variables that serve as the best predictors of an additive type. Figure 3.3 also shows a curve based on smoothing the data using this method. The smoothed curve shows an increasing trend. The trend seems approximately linear over most width values, and we next discuss models for which the mean or the log of the mean is linear in width.

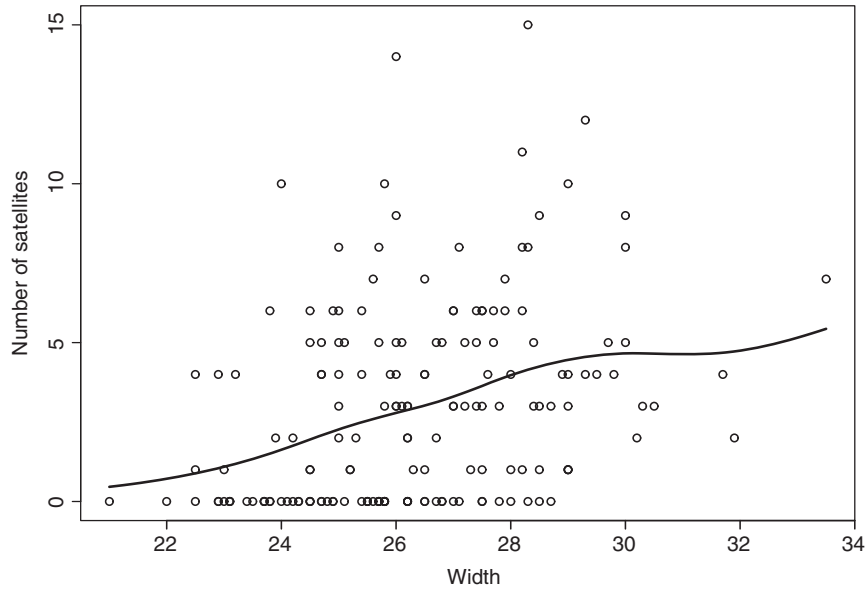


Figure 3.3 Number of satellites by female crab shell width (in centimeters), and generalized additive model smoothing fit.

For the expected number of satellites μ and shell width x for a female crab, GLM software reports the ML fit of the Poisson loglinear model as

$$\log \hat{\mu} = \hat{\alpha} + \hat{\beta}x = -3.305 + 0.164x.$$

Since $\hat{\beta} > 0$, width has a positive estimated effect on the number of satellites. Here is how to use the `glm` function in `R` to fit the model:

```
------------------------------------------------------------------------------
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
+                     header=TRUE)
> Crabs
    crab  sat   weight   width color spine # sat = number of satellites
1      1    8    3.050    28.3     2     3 # showing 3 of 173 observations
```

```
2      2    0    1.550    22.5     3     3
...
173   173    0    2.000    24.5     2     2
> plot(sat ~ width, xlab="Width", ylab="Number of satellites", data=Crabs)
> fit <- glm(sat ~ width, family=poisson(link=log), data=Crabs)
> # canonical link for Poisson is log, so "(link=log)" is not necessary
> summary(fit)
             Estimate  Std. Error
(Intercept)  -3.30476     0.54224
width         0.16405     0.01997
> library(gam) # generalized additive model smoothing fit
> gam.fit <- gam(sat ~ s(width), family=poisson, data=Crabs)
> # s() is smooth function predictor for generalized additive model
> curve(predict(gam.fit, data.frame(width=x), type="resp"), add=TRUE)
-------------------------------------------------------------------------
```

The model fit yields an estimated mean number of satellites $\hat{\mu}$, a *fitted value*, at any width. For instance, from (3.2), the fitted value at the mean width of $x = 26.3$ is

$$\hat{\mu} = \exp(\hat{\alpha} + \hat{\beta}x) = \exp[-3.305 + 0.164(26.3)] = 2.74.$$

For this model, $\exp(\hat{\beta}) = \exp(0.164) = 1.178$ represents the multiplicative effect on the fitted value for each 1-cm increase in x. For instance, the fitted value at $x = 27.3 = 26.3 + 1$ is $\exp[-3.305 + 0.164(27.3)] = 3.23$, which equals 1.178(2.74). A 1-cm increase in the shell width has an 17.8% increase in the estimated mean number of satellites.

The Poisson regression model with identity link function has ML fit

$$\hat{\mu} = -11.53 + 0.550x.$$

The effect of x on μ in this model is additive, rather than multiplicative. A 1-cm increase in shell width has an estimated increase of $\hat{\beta} = 0.55$ in the expected number of satellites. For instance, the fitted value at $\bar{x} = 26.3$ is $\hat{\mu} = -11.53 + 0.550(26.3) = 2.93$; at $x = 27.3$, it is $2.93 + 0.55 = 3.48$. The fitted values are positive at all observed sample widths, and the model provides a simple description of the width effect: on the average, a 2-cm increase in shell width corresponds to about an extra satellite.

Figure 3.4 plots the fitted number of satellites against width, for the models with a log link and with an identity link. Although they diverge somewhat for small and large width values, they provide similar predictions over the x-range in which most observations occur.

3.3.4 Overdispersion: Greater Variability than Expected

Count data often exhibit greater variation than we would expect if the response distribution truly were Poisson. The phenomenon of the data having greater variability than expected for a particular GLM is called *overdispersion*. A common cause of overdispersion is heterogeneity among subjects. For instance, suppose width, weight, color, and spine condition all affect $Y =$ number of satellites, which has a Poisson distribution at each fixed combination of those four variables. However, suppose the model uses width alone as a

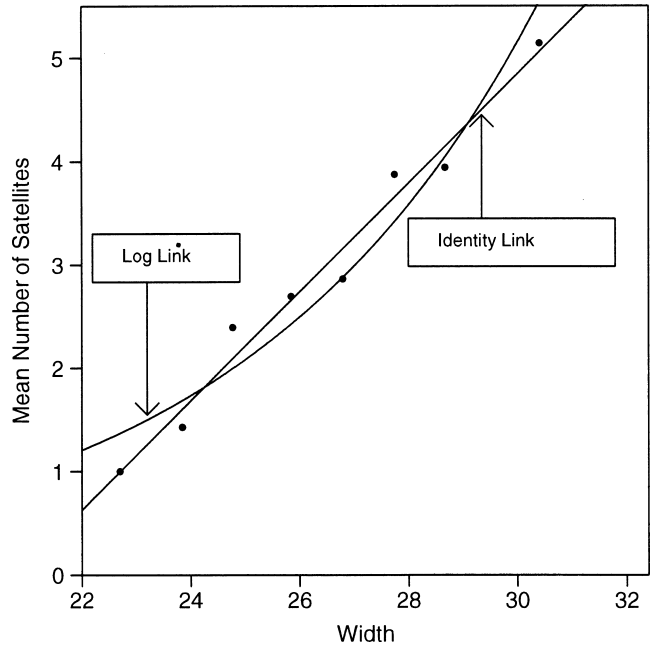


Figure 3.4 Estimated mean number of satellites for log and identity links.

predictor. Crabs having a certain width are a mixture of crabs of various weights, colors, and spine conditions. Thus, the population of crabs having that width is a mixture of several Poisson populations, each having its own mean for the response. This heterogeneity in the crabs having a certain width yields an overall response distribution at that width having greater variation than the Poisson predicts. If the variance equals the mean when we adjust for *all* relevant variables, it exceeds the mean when we adjust for only a *subset* of those variables.

Overdispersion is not an issue in ordinary regression models assuming normally distributed Y, because the normal has a separate parameter from the mean (i.e., the variance, σ^2) to describe variability. For Poisson distributions, however, the variance equals the mean. Overdispersion is common when we use Poisson GLMs for counts.

Chapter 7 presents Poisson GLMs for counts, with the main focus being the modeling of counts in contingency tables to investigate associations among categorical variables. We will learn about ways of handling overdispersion in Section 7.6. There we will find that the horseshoe crab satellite counts suffer from overdispersion and we will present a better way to model them.

3.4 STATISTICAL INFERENCE AND MODEL CHECKING

For GLMs, the ML estimators of model parameters have approximately normal sampling distributions for large probability samples. The three methods of inference introduced in Section 1.4 apply for any GLM. We review the methods here, in the context of GLMs with a single explanatory variable.

3.4.1 Wald, Likelihood-Ratio, and Score Inference Use the Likelihood Function

For a GLM with a single explanatory variable x, consider the hypothesis H_0: $\beta = 0$ that x has no effect on the response variable. The Wald test statistic is

$$z = \hat{\beta}/SE,$$

where *SE* is the unrestricted standard error of $\hat{\beta}$. It has an approximate standard normal distribution when H_0 is true. Equivalently, z^2 has an approximate chi-squared distribution with $df = 1$.

For the likelihood-ratio approach, denote the maximized value of the likelihood function by ℓ_0 under H_0: $\beta = 0$ and by ℓ_1 when β need not equal 0. The model with $\beta = 0$ has only an intercept term and is called the *null model*. The *likelihood-ratio* test statistic equals

$$2\log(\ell_1/\ell_0) = 2[\log(\ell_1) - \log(\ell_0)] = 2(L_1 - L_0),$$

where L_0 and L_1 denote the maximized log-likelihood functions. Under H_0: $\beta = 0$, this test statistic also has an approximate chi-squared distribution with $df = 1$.

Figure 3.5 shows a generic plot of a log-likelihood function $L(\beta)$ for a single parameter β, to illustrate tests of H_0: $\beta = 0$. The log-likelihood function for many GLMs, including binomial logistic regression models and Poisson loglinear models, has a concave (mound) shape such as Figure 3.5 shows. The ML estimate $\hat{\beta}$ is the point at which $L(\beta)$ takes its highest value. In Figure 3.5, the likelihood-ratio statistic $2(L_1 - L_0)$ is twice the vertical distance between values of $L(\beta)$ at $\hat{\beta}$ and at $\beta = 0$. The Wald test utilizes $L(\beta)$ only at the ML estimate $\hat{\beta}$. The SE of $\hat{\beta}$ is derived based on the curvature of $L(\beta)$ at $\hat{\beta}$. As n increases, the curvature of $L(\beta)$ increases (i.e., the log-likelihood becomes less

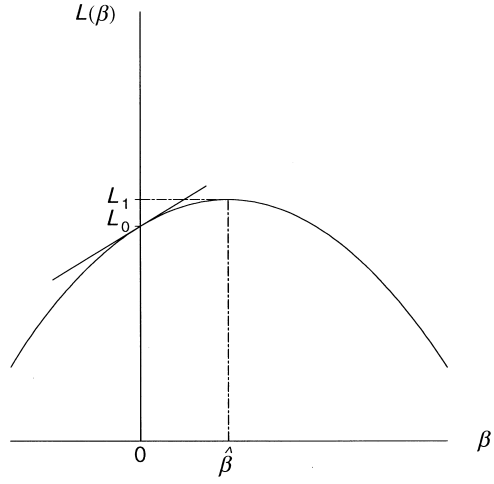


Figure 3.5 Information from the log-likelihood function $L(\beta)$ used in GLM tests of H_0: $\beta = 0$. The Wald test uses $\hat{\beta}$ and the curvature of $L(\beta)$ at $\hat{\beta}$. The likelihood-ratio test uses twice the difference $(L_1 - L_0)$ at $\beta = \hat{\beta}$ and at $\beta = 0$. The score test uses the slope of the line drawn tangent to $L(\beta)$ at $\beta = 0$.

flat), and standard errors become smaller. Greater curvature implies that the log-likelihood drops more quickly as β moves away from $\hat{\beta}$, and the range of plausible β values is then narrower.

The score test is based on the behavior of the log-likelihood function only at the null value for β of 0. It uses the magnitude of the slope of the line drawn tangent to $L(\beta)$ at $\beta = 0$. For a particular curvature, that slope tends to be larger in absolute value when $\hat{\beta}$ is farther from that null value. The z score statistic divides the slope by its SE, which is evaluated at the H_0 value[4] for β. Its square also has an approximate chi-squared null distribution with $df = 1$. We shall not present its general formula, but some commonly used test statistics for categorical data are of this type or generalizations of it for multiple parameters, such as the Pearson X^2 statistic and the ordinal correlation statistic M^2 for testing independence.

The three methods have corresponding confidence intervals. The 95% confidence interval for β consists of all β_0 values for which the P-value exceeds 0.05 in the test of H_0: $\beta = \beta_0$. For example, a Wald statistic for H_0 is $z = (\hat{\beta} - \beta_0)/SE$, which yields the Wald confidence interval $\hat{\beta} \pm z_{\alpha/2}(SE)$. The likelihood-ratio test-based confidence interval for β is called a *profile likelihood confidence interval*.

Which of the three methods should you prefer? Score tests and confidence intervals are not easily available for GLMs in most software. We will mainly use likelihood-ratio and Wald inference methods in this book. Wald inference is simple, similar to t methods for normal models, and directly available with any software that reports ML estimates and SE values. The likelihood-ratio inference is more cumbersome with some software. However, when n is small or effects are very large, it tends to be more powerful than Wald inference and is more trustworthy. Sometimes the Wald method fails completely.[5]

3.4.2 Example: Political Ideology and Belief in Evolution

The 2016 General Social Survey asked, "Human beings, as we know them today, developed from earlier species of animals. True or false?" Is the response to this question associated with one's political ideology? Let y = opinion about evolution (1 = true, 0 = false) and x = political ideology (1 = extremely conservative, 2 = conservative, 3 = slightly conservative, 4 = moderate, 5 = slightly liberal, 6 = liberal, 7 = extremely liberal). Here is edited R output for fitting and conducting inference for a logistic regression model, using the `Evolution` data file at the text website:

```
------------------------------------------------------------------------------
> Evo <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Evolution.dat",
+                   header=TRUE)
> Evo
  ideology true false
1        1   11    37 # extremely conservative
2        2   46   104
3        3   70    72
4        4  241   214
```

[4] Such as we did with formula (1.2) for the binomial parameter π in Section 1.3.2.

[5] Such as for inference about binomial parameters when $\hat{\pi} = 0$ or 1 or about any GLM parameter when $\hat{\beta} = \infty$ (e.g., Section 5.3).

```
5         5   78    36
6         6   89    24
7         7   36     6 # extremely liberal
> n <- Evo$true + Evo$false # binomial sample sizes
> fit <- glm(true/n ~ ideology, family=binomial, weights=n, data=Evo)
> summary(fit) # logistic regression fit
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -1.75658     0.20500   -8.569    <2e-16
ideology      0.49422     0.05092    9.706    <2e-16 # z Wald test
---
    Null deviance: 113.20 on 6 degrees of freedom # explained in Sec. 3.4.1
Residual deviance:   3.72 on 5 degrees of freedom
Number of Fisher Scoring iterations: 3 # explained in Section 3.5.1

> confint(fit) # profile likelihood CI
               2.5 %    97.5 %
ideology     0.39617   0.59594

> library(car)
> Anova(fit) # likelihood-ratio tests for effect parameters in a GLM
         LR Chisq  Df  Pr(>Chisq)
ideology   109.48   1   < 2.2e-16 # can also get with drop1(fit, test="LRT")
> library(statmod)
> fit0 <- glm(true/n ~ 1, family=binomial, weights=n, data=Evo) # null model
> glm.scoretest(fit0, Evo$ideology)^2 # squaring a z score statistic
[1] 104.101 # score chi-squared statistic with df=1
--------------------------------------------------------------------------------
```

The ML fit of the logistic regression model is $\text{logit}[\hat{P}(y = 1)] = -1.757 + 0.494x$. The ideology effect $\hat{\beta} = 0.494$ has $SE = 0.051$. The Wald test of H_0: $\beta = 0$ against H_a: $\beta \neq 0$ treats

$$z = \frac{\hat{\beta}}{SE} = \frac{0.494}{0.051} = 9.71$$

as standard normal under H_0, or $z^2 = 96.2$ as chi-squared with $df = 1$. This provides extremely strong evidence that belief in evolution increases as political ideology is more liberal ($P < 0.0001$). We obtain similar strong evidence from a likelihood-ratio test comparing this model to the simpler one having $\beta = 0$. In R, we can obtain this test by applying the `Anova` function from the `car` package. The chi-squared statistic equals $2(L_1 - L_0) = 109.48$ with $df = 1$ ($P < 0.0001$). The z score statistic is available with the `glm.scoretest` function in the `statmod` package, by adding the variable to be tested (here, ideology) to the model without it (here, the null model). Its square, which equals 104.10 for these data, is also a chi-squared statistic with $df = 1$ ($P < 0.0001$).

The Wald 95% confidence interval for β is $\hat{\beta} \pm 1.96(SE)$, which is $0.494 \pm 1.96(0.051)$, or (0.394, 0.594). From the R output, the profile likelihood 95% confidence interval for β is (0.396, 0.596).

3.4.3 The Deviance of a GLM

Let L_M denote the maximized log-likelihood value for a model M of interest. Let L_S denote the maximized log-likelihood value for the most complex model possible. This model has a separate parameter for each observation and it provides a perfect fit to the data. The model is said to be *saturated*.

For example, suppose M is the logistic regression model applied to the data just analyzed. The model for this 7×2 contingency table has 2 parameters for describing how the probability of belief in evolution changes for the 7 levels of $x =$ political ideology. The corresponding saturated model has a separate parameter for each of the 7 binomial observations: $P(Y = 1) = \pi_1$ for extremely conservative, π_2 for conservative, $\ldots$, π_7 for extremely liberal. The ML estimate for π_i in the saturated model is the sample proportion believing in evolution at level i of political ideology.

Because the saturated model has additional parameters, its maximized log-likelihood L_S is at least as large as the maximized log-likelihood L_M for a simpler model M. The *deviance* of a GLM is defined to be

$$\text{Deviance} = 2(L_S - L_M).$$

The deviance is the likelihood-ratio statistic for comparing model M to the saturated model. It is a test statistic for the hypothesis that all parameters that are in the saturated model but not in model M equal zero.

For some GLMs, the deviance has approximately a chi-squared distribution. For example, in Section 5.2.1 we will see this happens for binary GLMs with a fixed number of explanatory levels in which each observation is a binomial variate having relatively large counts of successes and failures. For such cases, the deviance provides a goodness-of-fit test of the model, because it tests the hypothesis that all possible parameters not included in the model equal 0. The residual *df* equals the number of observations minus the number of model parameters. The P-value is the right-tail probability above the observed test statistic value, from the chi-squared distribution. Large test statistics and small P-values provide strong evidence of model lack of fit.

When you fit a generalized linear model with the `glm` function in `R`, the reported *residual deviance* is the deviance for the model fitted, and the *null deviance* is the deviance for the null model, which has only the intercept term. For the evolution and political ideology data, the logistic regression model describes seven binomial observations by two parameters. The residual deviance, which the output in Section 3.4.2 reports to equal 3.72, has $df = 7 - 2 = 5$. For testing the null hypothesis that the model holds, the P-value is 0.59. The model seems to be adequate.

3.4.4 Model Comparison Using the Deviance

For any GLM, we can use the deviance to make inferential comparisons of two nested models M_0 and M_1, such that M_0 is a special case of M_1. For normal-response models, the F-test comparison of the models decomposes a sum of squares representing the variability in the data. This *analysis of variance* for decomposing variability generalizes to an *analysis of deviance* for GLMs. Given that the more complex model holds, the likelihood-ratio statistic for testing that the simpler model holds is $2(L_1 - L_0)$. Since

$$2(L_1 - L_0) = 2(L_S - L_0) - 2(L_S - L_1) = \text{Deviance}_0 - \text{Deviance}_1,$$

we can compare the models by comparing their deviances. This test statistic is large when M_0 fits poorly compared to M_1. For large samples, the statistic has an approximate chi-squared null distribution, with *df* equal to the difference between the residual *df* values for the separate models. This *df* value equals the number of additional parameters that are in M_1 but not in M_0. Large test statistics and small P-values suggest that M_1 fits better than M_0.

For the evolution and political ideology data, the output above in Section 3.4.2 shows that the deviance for the logistic regression model (i.e., the *residual* deviance) is 3.72 with $df = 5$. The simpler model with no effect of political ideology, which sets $\beta = 0$, has deviance (i.e., the *null* deviance) equal to 113.20 with $df = 6$. The difference between the deviances equals 109.48 with $df = 1$. However, this is precisely the likelihood-ratio statistic for testing H_0: $\beta = 0$ in the model. Generally, the difference between the null deviance and the residual deviance for a model is the likelihood-ratio statistic for testing that all the β effect terms in the model equal 0.

3.4.5 Residuals Comparing Observations to the Model Fit

For any GLM, goodness-of-fit statistics only broadly summarize how well models fit the data. We obtain further insight by comparing observed and fitted values individually.

For observation *i*, the difference $y_i - \hat{\mu}_i$ between an observed and fitted value has limited usefulness. For Poisson sampling, for instance, the standard deviation of a count is $\sqrt{\mu_i}$, so more variability tends to occur when μ_i is larger. The *Pearson residual* is a standardized difference:

$$\text{Pearson residual } = e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{var}(y_i)}}. \tag{3.3}$$

For Poisson GLMs, $\text{var}(y_i) = \mu_i$, so the Pearson residual for count i is

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

It divides the raw residual by the estimated Poisson standard deviation. The reason for calling e_i a *Pearson residual* is that $\sum e_i^2 = \sum_i (y_i - \hat{\mu}_i)^2/\hat{\mu}_i$. When the GLM is the model corresponding to independence for cells in a two-way contingency table, this is the Pearson chi-squared statistic X^2 for testing independence (equation (2.3)). Therefore, X^2 decomposes into terms describing the lack of fit for separate observations. *Deviance residuals* are alternative measures of lack of fit that decompose the deviance.

The denominator of the Pearson residual accounts for the variability in y_i but not for the variability in $\hat{\mu}_i$. The *standardized residual* (also called *standardized Pearson residual*) divides $(y_i - \hat{\mu}_i)$ by its estimated standard error,[6]

$$\text{Standardized residual} = \frac{y_i - \hat{\mu}_i}{SE}.$$

It accounts for both sources of variability and is preferable to the Pearson and deviance residuals. With standardized residuals, it is easier to tell when a deviation $(y_i - \hat{\mu}_i)$ is "large."

[6] $SE = \sqrt{\widehat{var}(y_i)(1 - h_i)}$, for *leverage* h_i. The formula for h_i is beyond our scope, but larger h_i indicate that y_i has more potential for influencing the model fit. Deviance residuals can also be standardized.

The standardized residuals have an approximate standard normal distribution when μ_i is large. Values larger than about 2 or 3 in absolute value are worthy of attention, although some values of this size occur by chance alone when the number of observations is large. Section 2.4.5 introduced standardized residuals that follow up tests of independence in two-way contingency tables. We will use standardized residuals with logistic regression in Section 5.2. Section 5.2.6 illustrates that they more appropriately reflect residual df than Pearson or deviance residuals.

The next output shows for each binomial in the `Evolution` data file: the political ideology, the *true* count for belief in evolution, the *false* count, the binomial sample size n, the sample proportion *true*, the logistic model fitted proportion, and the standardized residual. The sample proportions are close to the model fitted proportions. None of the standardized residuals indicate lack of fit, which is not surprising because the residual deviance suggested that the model fits well.

```
------------------------------------------------------------------------------
> attach(Evo)
> cbind(ideology,true,false,n,true/n,fitted(fit),rstandard(fit,type="pearson"))
>       # rstandard(fit, type="pearson") requests standardized residuals
 ideology  true false    n # sample  fitted  std. res.
1        1    11    37   48   0.2292  0.2206   0.1611 # extremely conservative
2        2    46   104  150   0.3067  0.3169  -0.3515
3        3    70    72  142   0.4930  0.4319   1.6480
4        4   241   214  455   0.5297  0.5549  -1.4995
5        5    78    36  114   0.6842  0.6714   0.3249
6        6    89    24  113   0.7876  0.7701   0.5414
7        7    36     6   42   0.8571  0.8459   0.2207 # extremely liberal
------------------------------------------------------------------------------
```

Other diagnostic tools from ordinary linear modeling are also helpful in assessing fits of GLMs. For instance, to assess the influence of an observation on the overall fit, you can refit the model with that observation deleted.

3.5 FITTING GENERALIZED LINEAR MODELS

In Statistics, maximizing the likelihood function to obtain ML estimates uses basic calculus. Taking derivatives of the log-likelihood function with respect to the various parameters and equating them to 0 yields "likelihood equations" that are solved to yield the ML estimates. Except in simple special cases, the likelihood equations for GLMs do not have closed-form solutions. Software uses an algorithm to solve them and fit the model.

3.5.1 The Fisher Scoring Algorithm Fits GLMs

An algorithm for fitting GLMs starts at an initial guess for the parameter values that maximize the likelihood function. Successive approximations produced by the algorithm tend to fall closer to the ML estimates. The *Fisher scoring* algorithm for doing this was first proposed by R.A. Fisher in 1935 for ML fitting of a binary regression model. Each cycle in the

algorithm represents a type of *weighted least squares* fitting. This is a generalization of ordinary least squares that accounts for nonconstant variance of Y in GLMs. Observations that occur where the variability is smaller receive greater weight in determining the parameter estimates. The weights change somewhat from cycle to cycle, with revised approximations for the ML estimates and variance estimates. ML estimation for GLMs is sometimes called *iteratively reweighted least squares*. The process is called *iterative* because the algorithm repeatedly uses the same type of step over and over until no further increase (in practical terms) occurs in the log-likelihood value. The successive approximations usually converge rapidly to the ML estimates, often within a few cycles. For example, the `R` code in Section 3.4.2 indicates that for the logistic regression model for political ideology and belief in evolution, Fisher scoring required only three iterations to converge.

For binomial logistic regression and Poisson loglinear models, Fisher scoring simplifies to a general-purpose method called the *Newton–Raphson algorithm*. That algorithm approximates the log-likelihood function in a neighborhood of the initial guess by a concave parabolic function that has the same slope and curvature at the initial guess as does the log-likelihood function. The location of the maximum of this approximating function comprises the second guess for the ML estimates. The algorithm then approximates the log-likelihood function in a neighborhood of the second guess by another concave parabolic function, and the third guess is the location of its maximum. The process iterates until convergence.

3.5.2 Bayesian Methods for Generalized Linear Models

Effect parameters $\{\beta_j\}$ in generalized linear models can take value over the entire real line. In Bayesian inference about $\{\beta_j\}$, considerable flexibility for their prior distribution is provided by the family of multivariate normal distributions. A simple, relatively uninformative, prior distribution has an independent normal distribution for each model parameter, with a mean of 0 and a large standard deviation.

When we use the same standard deviation for an uninformative prior distribution for each model parameter, it is sensible to standardize quantitative explanatory variables (i.e., to have means of 0 and standard deviations of 1) so that the prior effects are identical in size. Otherwise, take the scale into account. For example, if $x_j =$ time is rescaled from years to months, the new β_j is 1/12th as large, so you should multiply σ in the normal prior by 1/12 compared to when x is measured in years.

Software using Monte Carlo simulation methods is now available that makes Bayesian methods computationally feasible for GLMs. We illustrate in Section 5.4 for logistic regression models and in Section 7.2.5 for loglinear models.

3.5.3 GLMs: A Unified Approach to Statistical Analysis

The development of GLM theory in the mid-1970s unified important models for continuous and discrete response variables. Table 3.3 lists several popular GLMs for statistical analyses.

A pleasing feature of GLMs is that the model-fitting algorithm, Fisher scoring, is the same for any GLM. This holds regardless of the choice of the distribution for Y or the link function. GLM software can therefore fit a very wide variety of useful models.

Table 3.3 Generalized linear models for statistical analysis.

Random Component	Link Function	Explanatory Variables	Model	Chapter
Normal	Identity	Continuous	Regression	
Normal	Identity	Categorical	Analysis of variance	
Normal	Identity	Mixed	Analysis of covariance	
Binomial	Logit	Mixed	Logistic regression	4–5, 8–10
Multinomial	Logits	Mixed	Multinomial logit	6, 8–10
Poisson	Log	Mixed	Loglinear	7

EXERCISES

3.1 Describe the purpose of the link function of a GLM. Define the identity link and explain why it is not often used with a binomial parameter.

3.2 In the years 1904, 1914, 1924, . . . , 2014, the percentage of times the starting pitcher pitched a complete game were[7]: 87.6, 55.0, 48.7, 43.4, 45.2, 34.0, 24.5, 28.0, 15.0, 8.0, 3.1, 2.4.

a. The linear probability model has least squares fit $\hat{P}(Y = 1) = 0.6930 - 0.0662x$, where $x =$ number of decades since 1904. Interpret -0.0662.

b. Substituting $x = 12$ in the linear prediction equation, predict the proportion of complete games for 2024. The ML fit of the logistic regression model yields $\hat{P}(Y = 1) = 0.034$ at $x = 12$. Which prediction is more plausible? Why?

3.3 For Table 2.6 on $x =$ mother's alcohol consumption and $Y =$ whether a baby has sex organ malformation, ML fitting of the linear probability model with x scores 0, 0.5, 1.5, 4.0, 7.0 has output:

```
------------------------------------------
Parameter   Estimate  Std Error
Intercept   0.00255    0.0003
alcohol     0.00109    0.0007
------------------------------------------
```

a. State the prediction equation and interpret the intercept and slope.

b. Use the model fit to estimate the (i) probabilities of malformation for alcohol levels 0 and 7.0, (ii) relative risk comparing those levels.

c. Is the result sensitive to the choice of scores? Re-fit the linear probability model using scores 0, 1, 2, 3, 4, and re-evaluate fitted probabilities at alcohol levels 0 and 7 and the relative risk.

d. The sample proportion of malformations is much higher in the highest alcohol category than the others because, although it has only one malformation, its sample size is only 38. Are results sensitive to this single malformation? Fit the logistic regression or linear probability model with and without that observation, and evaluate fitted probabilities at alcohol levels 0 and 7 and the relative risk.

[7] *Source*: `https://en.wikipedia.org/wiki/Complete_game`.

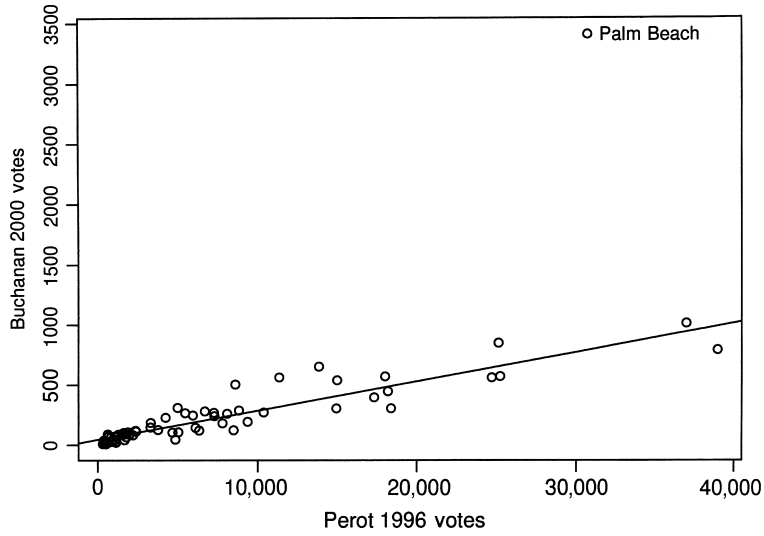


Figure 3.6 Total vote, by county in Florida, for Reform Party candidates Buchanan in 2000 and Perot in 1996.

3.4 In the 2000 U.S. Presidential election, Palm Beach County in Florida was the focus of unusual voting patterns apparently caused by a confusing "butterfly ballot." Many voters claimed they voted mistakenly for the Reform Party candidate, Pat Buchanan, when they intended to vote for Al Gore. Figure 3.6 shows the total number of votes for Buchanan plotted against the number of votes for the Reform Party candidate in 1996 (Ross Perot), by county in Florida.[8]

a. In county i, let π_i denote the proportion of the vote for Buchanan and let x_i denote the proportion of the vote for Perot in 1996. For the linear probability model fitted to all counties except Palm Beach County, $\hat{\pi}_i = -0.0003 + 0.0304x_i$. Give the value of P in the interpretation. The estimated proportion vote for Buchanan in 2000 was roughly P% of that for Perot in 1996.

b. For Palm Beach County, $\pi_i = 0.0079$ and $x_i = 0.0774$. Does this result appear to be an outlier for the model? Investigate, by finding $\pi_i/\hat{\pi}_i$ and $\pi_i - \hat{\pi}_i$. (Statistical analyses predicted that fewer than 900 votes were truly intended for Buchanan, compared to the 3407 he received. George W. Bush won the state by 537 votes and, with it, the Electoral College and the election. Other ballot design problems played a role in 110,000 disqualified overvote ballots, in which people mistakenly voted for more than one candidate, with Gore marked on 84,197 ballots and Bush on 37,731.)

3.5 Access the horseshoe crab data file (shown partly in Table 3.2) at `www.stat.ufl.edu/~aa/cat/data`. In the data file, $y = 1$ if a crab has at least one satellite and $y = 0$ otherwise.

a. Using weight as the predictor, fit the linear probability model to $P(Y = 1)$. If your software cannot use the identity link with the binomial or fails to converge, use

[8] For details, see A. Agresti and B. Presnell, *Statistical Science* **17**: 436–440 (2002).

ordinary least squares by treating Y as normally distributed. Interpret the parameter estimates. Find $\hat{P}(Y = 1)$ at the highest observed weight of 5.20 kg. Comment.

b. Fit the logistic regression model. Show that at a weight of 5.20 kg, $\hat{P}(Y = 1) = 0.9968$.

3.6 From the 2016 General Social Survey, when we cross-classify political ideology (with 1 being most liberal and 7 being most conservative) by political party affiliation for subjects of ages 18–27, we get:

```
-------------------------------------------------------------
               1    2    3    4    5    6    7
Democrat       5   18   19   25    7    7    2
Republican     1    3    1   11   10   11    1
-------------------------------------------------------------
```

When we use R to model the effect of political ideology on the probability of being a Democrat, we get the results:

```
-------------------------------------------------------------
> y <- c(5,18,19,25,7,7,2); n <- c(6,21,20,36,17,18,3)
> x <- c(1,2,3,4,5,6,7)
> fit <- glm(y/n ~ x, family=binomial(link=logit), weights=n)
> summary(fit)
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  3.1870      0.7002    4.552  5.33e-06
x           -0.5901      0.1564   -3.772  0.000162
---
    Null deviance: 24.7983  on 6  degrees of freedom
Residual deviance:  7.7894  on 5  degrees of freedom
Number of Fisher Scoring iterations: 4
> confint(fit)
                2.5 %    97.5 %
(Intercept)   1.90180   4.66484
x            -0.91587  -0.29832
-------------------------------------------------------------
```

a. Report the prediction equation and interpret the direction of the estimated effect.

b. Construct the 95% Wald confidence interval for the effect of political ideology. Interpret and compare to the profile likelihood interval shown.

c. Conduct the Wald test for the effect of x. Report the test statistic, P-value, and interpret.

d. Conduct the likelihood-ratio test for the effect of x. Report the test statistic, find the P-value, and interpret.

e. Explain the output about the number of Fisher scoring iterations.

3.7 Consider Table 3.1 on snoring and heart disease.

a. Re-fit the logistic regression model using the scores (i) (0, 2, 4, 6), (ii) (0, 1, 2, 3), (iii). (1, 2, 3, 4). Compare the model parameter estimates under the three choices.

Compare the fitted values. What can you conclude about the effect of *linear* transformations of scores that preserve relative sizes of spacings between scores?

b. Fit the logistic regression model using the scores (0, 2, 6, 7), approximating the number of days in a week that the subject snores. Compare fitted values to those with the scores (0, 2, 4, 5) used in the text example. Do results seem to be sensitive to the choice of scores?

3.8 Fit the logistic regression model of Section 3.2.3 to Table 3.1 on snoring and heart disease. Show results of significance tests and confidence intervals for the effect of snoring.

3.9 Table 3.4, the `Credit` data file at the text website, shows data for a sample of 100 adults randomly selected for an Italian study on the relation between $x =$ annual income and $y =$ whether you have a travel credit card ($1 =$ yes, $0 =$ no). At each level of x (in thousands of euros), the table indicates the number of subjects in the sample and the number of those having at least one travel credit card. Software provides the following results of using logistic regression:

```
---------------------------------------------
             Estimate  Std. Error
(Intercept)  -3.5179     0.7103
x             0.1054     0.0262
---------------------------------------------
```

a. Report the prediction equation and interpret the sign of $\hat{\beta}$.

b. When $\hat{P}(Y = 1) = 0.50$, show that the estimated logit value is 0. Based on this, for these data explain why the estimated probability of a travel credit card is 0.50 at income $= 33.4$ thousand euros.

c. Show how to apply software to the `Credit` data file at the text website to obtain the logistic regression fit.

Table 3.4 Data on travel credit cards and income for exercise 3.9.

Income	No. of Cases	Credit Cards	Income	No. of Cases	Credit Cards	Income	No. of Cases	Credit Cards
12	1	0	21	2	0	34	3	3
13	1	0	22	1	1	35	5	3
14	8	2	24	2	0	39	1	0
15	14	2	25	10	2	40	1	0
16	9	0	26	1	0	42	1	0
17	8	2	29	1	0	47	1	0
19	5	1	30	5	2	60	6	6
20	7	0	32	6	6	65	1	1

Source: Thanks to R. Piccarreta, Bocconi University, Milan, for original form of data.

3.10 A recent General Social Survey asked "How many people at your work place are close friends?" The 756 responses had a mean of 2.76, standard deviation of 3.65, and a mode of 0. Would the Poisson distribution describe these data well? Why or why not?

3.11 An experiment analyzes imperfection rates for two processes used to fabricate silicon wafers for computer chips. For treatment A applied to 10 wafers, the numbers of imperfections are 8, 7, 6, 6, 3, 4, 7, 2, 3, 4. Treatment B applied to 10 other wafers has 9, 9, 8, 14, 8, 13, 11, 5, 7, 6 imperfections. Treat the counts as independent Poisson variates having means $\mu_{\rm A}$ and $\mu_{\rm B}$. Consider the model $\log \mu = \alpha + \beta x$, where $x = 1$ for treatment B and $x = 0$ for treatment A, for which $\beta = \log \mu_{\rm B} - \log \mu_{\rm A} = \log(\mu_{\rm B}/\mu_{\rm A})$ and $e^{\beta} = \mu_B/\mu_{\rm A}$. Fit the model. Report the prediction equation and interpret $\hat{\beta}$.

3.12 Refer to the previous exercise.

a. Test H_0: $\mu_{\rm A} = \mu_{\rm B}$ by conducting the Wald or likelihood-ratio test of H_0: $\beta = 0$. Interpret.

b. Construct a 95% confidence interval for $\mu_{\rm B}/\mu_{\rm A}$. (*Hint*: Construct one for $\beta = \log(\mu_{\rm B}/\mu_{\rm A})$ and then exponentiate.)

3.13 For the `Crabs` data file (partially shown in Table 3.2) at `www.stat.ufl.edu/~aa/cat/data`, fit the Poisson loglinear model to use weight to predict the number of satellites.

a. Report the prediction equation, and estimate the mean response for female crabs of average weight, 2.44 kg.

b. Use $\hat{\beta}$ to describe the weight effect. Construct a 95% confidence interval for β and for the multiplicative effect of a 1-kg increase.

c. Conduct Wald and likelihood-ratio tests of the hypothesis that the mean response is independent of weight. Interpret.

3.14 If you are modeling count data, explain why it is not sufficient to analyze ordinary raw residuals, $(y_i - \hat{\mu}_i)$, as you would for ordinary linear models.

3.15 True or false?

a. An ordinary regression model that treats Y as normally distributed is a special case of a GLM, with a normal random component and identity link function.

b. With a GLM, Y does not need to have a normal distribution and one can model a function of the mean of Y instead of just the mean itself, but to get ML estimates the variance of Y must be constant at all values of explanatory variables.

CHAPTER 4

LOGISTIC REGRESSION

We now focus on the statistical modeling of binary response variables, for which the response outcome for each subject is a *success* or a *failure*. Binary data are the most common form of categorical data. The methods of this chapter are of fundamental importance, because binary data are common but also because the methods extend directly to nominal and ordinal response variables.

The most popular model for binary data is *logistic regression*. Section 3.2.2 introduced this model as a generalized linear model (GLM) for a binomial random component with the logit link function. We will now present ways to interpret the model, to conduct statistical inference for its parameters, and to summarize effects and predictive power. We will start with a single explanatory variable and then present extensions for multiple explanatory variables, which can be both quantitative and categorical.

4.1 THE LOGISTIC REGRESSION MODEL

For a binary response variable Y, we model the *success* probability, $P(Y = 1)$. In the single explanatory variable case, we denote this by $\pi(x)$ to emphasize that the value of $P(Y = 1)$ depends on the value x of that variable. We assume that the observations are independent binomial variates with parameter $\pi(x)$, which itself varies according to the value of x.

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

4.1.1 The Logistic Regression Model

The logistic regression model[1] has a linear form for the *logit* of the success probability, that is, the logarithm of the odds,

$$\text{logit}[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x. \tag{4.1}$$

For quantitative x, the formula implies that $\pi(x)$ changes as an S-shaped function of x, as shown in Section 3.2.2. Logistic regression has a corresponding formula for $\pi(x)$, using the exponential function $\exp(\alpha+\beta x) = e^{\alpha+\beta x}$,

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}. \tag{4.2}$$

The effect parameter β determines the rate of increase or decrease of the S-shaped curve for $\pi(x)$. The sign of β indicates whether the curve ascends ($\beta > 0$) or descends ($\beta < 0$). The rate of change increases as $|\beta|$ increases. When $\beta = 0$, the curve flattens to a horizontal straight line. The binary response variable is then independent of the explanatory variable.

4.1.2 Odds Ratio and Linear Approximation Interpretations

The logistic regression formula (4.1) indicates that the logit increases by β for every 1-unit increase in x. Most of us do not think naturally on a logit scale, so we next suggest alternative interpretations.

By exponentiating both sides of the logistic regression equation (4.1), we obtain an interpretion that uses the *odds* and the *odds ratio*. The odds of a success are

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\alpha+\beta x) = e^{\alpha}(e^{\beta})^{x}.$$

Therefore, the odds multiply by e^{β} for every 1-unit increase in x. That is, the odds at level $x+1$ equal the odds at x multiplied by e^{β}. When $\beta = 0$, $e^{\beta} = 1$, and the odds do not change as x changes.

A simpler interpretation refers to the probability $\pi(x)$ itself. Figure 4.1 shows the S-shaped appearance of the model for $\pi(x)$, as fitted for the next example. Since it is curved rather than a straight line, the rate of change in $\pi(x)$ per 1-unit increase in x depends on the value of x. A straight line drawn tangent to the curve at a particular x value, such as shown in Figure 4.1, describes the rate of change at that point. For logistic regression parameter β, that line has slope equal to $\beta\pi(x)[1-\pi(x)]$. For instance, the line tangent to the curve at x for which $\pi(x) = 0.50$ has slope $\beta(0.50)(0.50) = 0.25\beta$; by contrast, when $\pi(x) = 0.90$ or 0.10, it has slope 0.09β. The slope approaches 0 as $\pi(x)$ approaches 1.0 or 0. The steepest slope occurs when $\pi(x) = 0.50$. That x value relates to the logistic regression parameters by[2] $x = -\alpha/\beta$. This x value is sometimes called the *median effective level*. It represents the point at which each outcome has a 50% chance.

[1] First proposed in 1944 by the statistician and physician Joseph Berkson.

[2] You can show this by substituting $\pi(x) = 0.50$ in equation (4.1) and solving for x.

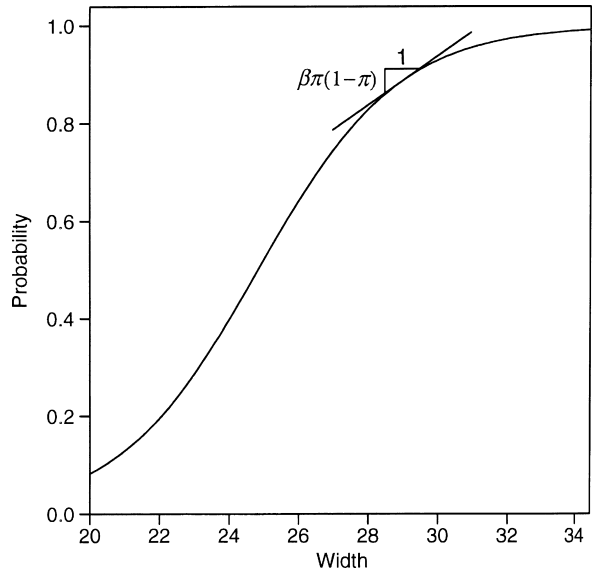


Figure 4.1 Linear approximation to the logistic regression curve.

4.1.3 Example: Whether a Female Horseshoe Crab Has Satellites

We now further analyze the horseshoe crab data introduced in Section 3.3.3 and available in the `Crabs` data file at the text website. Here, we let y indicate whether a female crab has any satellites (other males who could mate with her). That is, $y = 1$ if a female crab has at least one satellite and $y = 0$ if she has no satellite. We first use the female crab's shell width, in centimeters (cm), as the sole explanatory variable. It takes values between 21.0 and 33.5 cm.

Figure 4.2 plots the data. The plot consists of a set of points at the level $y = 1$ and a second set of points at the level $y = 0$. This figure jitters the data points slightly in the y direction so we can see coordinates at which multiple observations occur. It appears that $y = 1$ occurs relatively more often at higher x values. Since y takes only values 0 and 1, however, it is difficult to determine from a scatterplot whether a logistic regression model is appropriate.

Software can smooth the data to suggest the form of the relationship. Figure 4.2 also shows a curve that is the fit of a *generalized additive model*, introduced in Section 11.4, which allows the effect of x to have an arbitrary nonlinear form. This smoothing curve shows an increasing trend, so we proceed with fitting a model that implies such a trend.

Here is `R` code and edited output for fitting generalized additive and logistic regression models and plotting the data and the model fits:

```
------------------------------------------------------------------------------
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
+                     header=TRUE)
> Crabs
    crab  sat   y  weight   width color spine
1      1    8   1   3.050    28.3     2     3 # showing 3 of 173 observations
2      2    0   0   1.550    22.5     3     3 # y is satellite indicator
```

```
...                                          # y = 1 when sat > 0
173  173    0   0   2.000    24.5     2     2
> plot(jitter(y, 0.08) ~  width, data=Crabs) # scatterplot of y by x=width
> library(gam) # for fitting generalized additive models
> gam.fit <- gam(y ~ s(width), family=binomial, data=Crabs) # s = smooth funct.
> curve(predict(gam.fit, data.frame(width=x), type="resp"), add=TRUE)
>       # plots generalized additive model smoothing fit
> fit <- glm(y ~ width, family=binomial, data=Crabs) # link=logit is default
> curve(predict(fit, data.frame(width=x), type="resp"), add=TRUE)
>       # logistic regression fit is added to the plot
> summary(fit)
              Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -12.3508      2.6287   -4.698  2.62e-06
width           0.4972      0.1017    4.887  1.02e-06 # estimated beta = 0.4972

> predict(fit, data.frame(width = 21.0), type="response")
0.12910 # estimated probability of satellite at width = 21.0
> predict(fit, data.frame(width = mean(Crabs$width)), type="response")
0.67388 # estimated probability of satellite at mean width
-------------------------------------------------------------------------------
```

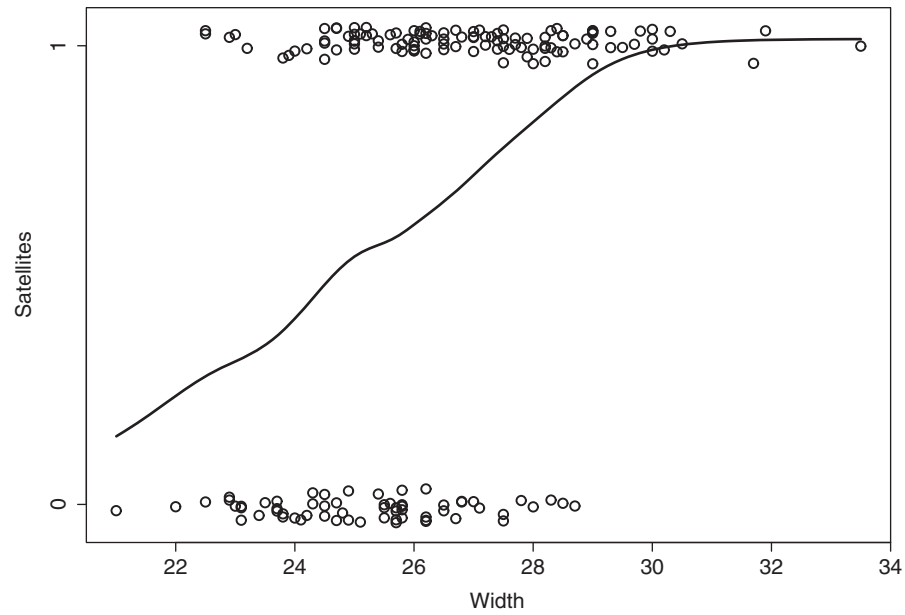


Figure 4.2 Whether horseshoe crab satellites are present ($y = 1$, yes; $y = 0$, no), by $x =$ width of shell, and generalized additive model smoothing fit.

For probability $\pi(x)$ that a female horseshoe crab of width x has a satellite, the logistic regression model fit is

$$\text{logit}[\hat{\pi}(x)] = -12.351 + 0.497x.$$

The estimated probability of a satellite is the sample analog of formula (4.2),

$$\hat{\pi}(x) = \frac{\exp(-12.351 + 0.497x)}{1 + \exp(-12.351 + 0.497x)}.$$

Since $\hat{\beta} > 0$, the estimated probability is higher at larger width values. At the minimum width in this sample of 21.0 cm, the estimated probability that the crab has at least one satellite is

$$\frac{\exp[-12.351 + 0.497(21.0)]}{1 + \exp[-12.351 + 0.497(21.0)]} = 0.129.$$

At the maximum sample width of 33.5 cm, the estimated probability equals 0.987. The effect of width seems relatively strong, in the sense that $\hat{\pi}(x)$ changes substantially over the range of x values. The median effective level, at which $\hat{\pi}(x) = 0.50$, is $x = -\hat{\alpha}/\hat{\beta} = 12.351/0.497 = 24.8$. Figure 4.1 plots the estimated probabilities.

At the sample mean width of 26.3 cm, $\hat{\pi}(x) = 0.674$. From Section 4.1.2, the incremental rate of change in the fitted probability at that point is $\hat{\beta}\hat{\pi}(x)[1 - \hat{\pi}(x)] = 0.497(0.674)(0.326) = 0.11$. For female crabs near the mean width, the estimated probability of having at least one satellite increases at the rate of 0.11 per 1-cm increase in width.

Since $\hat{\beta} = 0.497$, the estimated odds of a satellite multiply by $\exp(\hat{\beta}) = \exp(0.497) = 1.64$ for each 1-cm increase in width; that is, there is a 64% increase. To illustrate, the mean width value of $x = 26.3$ has $\hat{\pi}(x) = 0.674$, and odds = 0.674/0.326 = 2.07. At $x = 27.3 = 26.3 + 1.0$, you can check that $\hat{\pi}(x) = 0.773$ and odds = 0.773/0.227 = 3.40. However, this is a 64% increase; that is, $3.40 = 2.07(1.64)$.

4.1.4 Logistic Regression with Retrospective Studies

Another property of logistic regression relates to situations in which the explanatory variable X rather than the response variable Y is random. This occurs with retrospective sampling designs, such as in biomedical case-control studies (Section 2.3.5). For samples of subjects having $y = 1$ (cases) and having $y = 0$ (controls), the value of X is observed. Evidence exists of an association between X and Y if the distribution of X values differs between cases and controls. For binary X, we then treat the observations on X, given y, as binomial samples, instead of the observations on Y, given x. The odds ratio for X given y equals that for Y given x, so we can estimate odds ratios in such studies. Logistic regression effect parameters refer to odds and odds ratios, so we can estimate them.

For example, Table 2.3 in Section 2.3.5 showed results of a case-control study that matched 709 lung cancer cases with 709 controls and observed the association with smoking. We can regard the odds ratio of 3.0 for that table as an estimate of e^{β} for the model

$$\text{logit}[P(Y = 1)] = \alpha + \beta x,$$

where y is the lung cancer indicator and $x = 1$ for smokers and $x = 0$ for nonsmokers. We cannot estimate the intercept term α in the model, because it relates[3] only to the distribution of Y given x rather than X given y.

[3] For example, $\alpha = \text{logit}[P(Y = 1)]$ for nonsmokers and $\alpha + \beta = \text{logit}[P(Y = 1)]$ for smokers.

With case-control studies, it is not possible to estimate effects in binary models with link functions other than the logit. The effect measure is not then an odds ratio, so the effect for the conditional distribution of X given y does not then equal that for Y given x. This provides an important advantage of the logit link. It is a major reason why logistic regression surpasses other binary regression models in popularity for biomedical research.[4]

4.1.5 Normally Distributed *X* Implies Logistic Regression for *Y*

Regardless of the sampling mechanism, the logistic regression model may or may not describe a relationship well. In one special case, it does necessarily hold. Suppose the distribution of X is $N(\mu_1, \sigma^2)$ for subjects having $y = 1$ and $N(\mu_0, \sigma^2)$ for subjects having $y = 0$; that is, both normal, with possibly different means but with the same variance. Then, a Bayes theorem calculation converting from the distribution of X given y to the distribution of Y given x shows that $P(Y = 1 \mid x)$ satisfies the logistic regression curve. For that curve, the effect of x is $\beta = (\mu_1 - \mu_0)/\sigma^2$. In particular, β has the same sign as $\mu_1 - \mu_0$. If those with $y = 1$ tend to have higher values of x, then $\beta > 0$.

For example, consider $Y =$ heart disease ($1 =$ yes, $0 =$ no) and $X =$ cholesterol level. Suppose cholesterol levels have approximately a normal distribution when $y = 1$ and when $y = 0$, with $\mu_0 = 160$, $\mu_1 = 260$, and $\sigma = 50$. Then, the probability of having heart disease satisfies the logistic regression function (4.2) with predictor x and $\beta = (260 - 160)/50^2 = 0.04$.

If the distributions of X are bell-shaped but with highly different spreads when $y = 1$ and when $y = 0$, then a logistic model containing also a quadratic term (i.e., both x and x^2) often fits well. In that case, the relationship is not monotone. Instead, $P(Y = 1)$ increases and then decreases, or the reverse (Exercise 4.7).

4.2 STATISTICAL INFERENCE FOR LOGISTIC REGRESSION

Statistical inference for the logistic regression model parameters helps us judge the significance and size of the effects of explanatory variables. We use the Wald and likelihood-ratio methods introduced in Section 1.4.1 and presented for GLMs in Section 3.4.1.

As in other statistical analyses, standard errors tend to decrease as the total sample size n increases. For fixed n, however, the SE values are relatively large when the estimated probabilities $\{\hat{\pi}_i\}$ are mainly close to 0 or close to 1. For example, it is more difficult to estimate effects of explanatory variables well when nearly all the observations are successes (e.g., as in modeling the occurrence of a rare disease) compared to when a similar number of successes and failures occurs.

4.2.1 Confidence Intervals for Effects

A Wald confidence interval for the effect in the logistic regression model, $\text{logit}[\pi(x)] = \alpha + \beta x$, is $\hat{\beta} \pm z_{\alpha/2}(SE)$. Exponentiating the endpoints yields an interval for e^{β}, the multiplicative effect on the odds of a 1-unit increase in x. When n is small or when fitted

[4] Section 8.2.5 presents specialized methods for employing logistic regression with matched case-control studies.

probabilities are mainly near 0 or 1, it is preferable to construct a profile likelihood confidence interval.

For logistic regression modeling of the horseshoe crab data in terms of the probability that a female crab has at least one male satellite, with shell width as the explanatory variable (Section 4.1.3), the following edited R output shows inferences about the width effect:

```
------------------------------------------------------------------------------
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
+                       header=TRUE)
> fit <- glm(y ~ width, family=binomial, data=Crabs)
> summary(fit)
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -12.3508      2.6287   -4.698  2.62e-06
width         0.4972      0.1017    4.887  1.02e-06 # z Wald test
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of freedom

> confint(fit) # profile likelihood confidence interval
                  2.5 %    97.5 %
width           0.30838   0.70902

> library(car)
> Anova(fit) # likelihood-ratio test of width effect
      LR Chisq  Df  Pr(>Chisq)  # also shown with drop1(fit, test="LRT")
width   31.306   1   2.204e-08
------------------------------------------------------------------------------
```

Since $\hat{\beta} = 0.497$ with $SE = 0.102$, the 95% Wald confidence interval for β is $0.497 \pm 1.96(0.102)$, or (0.298, 0.697). The profile likelihood confidence interval is (0.308, 0.709), for which the interval for the effect on the odds is $(e^{0.308}, e^{0.709}) = (1.36, 2.03)$. We infer that a 1-cm increase in width has at least a 36% increase and at most a doubling in the odds that a female crab has a satellite.

From Section 4.1.2, a simpler interpretation uses the straight-line approximation $\beta\pi(x)[1 - \pi(x)]$ for the change in the probability per 1-unit increase in x. For instance, at $\pi(x) = 0.50$, the estimated rate of change is $0.25\hat{\beta} = 0.124$. A 95% confidence interval for 0.25β equals 0.25 times the endpoints of the interval for β. For the profile likelihood interval, this is [0.25(0.308), 0.25(0.709)] = (0.077, 0.177). Therefore, if the logistic regression model holds, then for values of x near the width value at which $\pi(x) = 0.50$, we infer that the rate of increase in the probability of a satellite per 1-cm increase in width falls between about 0.08 and 0.18.

4.2.2 Significance Testing

For the logistic regression model, H_0: $\beta = 0$ states that the probability of success is independent of x. For large samples, the Wald test statistic $z = \hat{\beta}/SE$ has a standard normal distribution when $\beta = 0$. Equivalently, for the two-sided H_a: $\beta \neq 0$, $z^2 = (\hat{\beta}/SE)^2$ has a large-sample chi-squared null distribution with $df = 1$. The likelihood-ratio test is more

powerful and more reliable than the Wald test, especially when n is small or when the effect β is strong. Let L_0 denote the maximum of the log-likelihood function when $\beta = 0$, which is the *null model*, containing only an intercept term. Let L_1 denote the maximum log-likelihood for unrestricted β. The test statistic, $2(L_1 - L_0)$, also has a large-sample chi-squared null distribution with $df = 1$. This statistic equals the difference between the null deviance for the null model and the residual deviance for the model containing the explanatory variable.

For the output in Section 4.2.1, the Wald statistic $z = \hat{\beta}/SE = 0.497/0.102 = 4.89$ shows strong evidence of a positive effect of width on the presence of satellites ($P < 0.0001$). The equivalent chi-squared statistic, $z^2 = 23.88$, has $df = 1$. Since the null deviance is 225.76 ($df = 172$) and the residual deviance is 194.45 ($df = 171$), the likelihood-ratio statistic for testing H_0: $\beta = 0$ is $225.76 - 194.45 = 31.31$, with $df = 172 - 171 = 1$. This also provides extremely strong evidence of a width effect. We can obtain this test in R directly using the `Anova` function in the `car` package.

4.2.3 Fitted Values and Confidence Intervals for Probabilities

Software for logistic regression can report the estimate of $P(Y = 1)$,

$$\hat{P}(Y = 1) = \exp(\hat{\alpha} + \hat{\beta}x)/[1 + \exp(\hat{\alpha} + \hat{\beta}x)],$$

for each observation, which is a *fitted value*. It can also construct a confidence interval for $P(Y = 1)$ at each x value, under the assumption that the model holds. If your software does not provide an option for the confidence interval, you can obtain it by finding a confidence interval for the linear predictor value[5] $\alpha + \beta x$ and then applying the $\exp()/[1 + \exp()]$ transform to the endpoints.

For female crabs of width $x = 26.5$, which is near the mean width, the estimated probability of a satellite is $\hat{P}(Y = 1) = 0.695$ and a 95% confidence interval for $P(Y = 1)$ is (0.61, 0.77). Here is how to use R to find such point and interval estimates at all observed values of the explanatory variables:

```
------------------------------------------------------------------------------
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
+                     header=TRUE)
> fit <- glm(y ~ width, family=binomial, data=Crabs)
> pred.prob <- fitted(fit) # ML fitted value estimate of P(Y=1)
> lp <- predict(fit, se.fit=TRUE) # linear predictor
> LB <- lp$fit - 1.96*lp$se.fit # confidence bounds for linear predictor
> UB <- lp$fit + 1.96*lp$se.fit # better: use qnorm(0.975) instead of 1.96
> LB.p <- exp(LB)/(1 + exp(LB)) # confidence bounds for P(Y=1)
> UB.p <- exp(UB)/(1 + exp(UB))
> cbind(Crabs$width, pred.prob, LB.p, UB.p)
          pred.prob     LB.p     UB.p
1    28.3   0.84823  0.75285  0.91114
```

[5] For example, using the covariance matrix of the ML parameter estimates, $(\hat{\alpha} + \hat{\beta}x) \pm 1.96\sqrt{var(\hat{\alpha}) + x^2 var(\hat{\beta}) + 2x[cov(\hat{\alpha}, \hat{\beta})]}$.

```
...
7    26.5   0.69546  0.61205  0.76775 # confidence bounds when width = 26.5
...
173  24.5   0.45793  0.35552  0.56403
-------------------------------------------------------------------------------
```

Figure 4.3 shows the estimated probabilities and the lower and upper 95% confidence bands. To obtain this figure after fitting the model, you can use the following code:

```
-------------------------------------------------------------------------------
> plot(jitter(y,0.1) ~ width, xlim=c(18,34), pch=16, ylab="Prob(satellite)",
+      data=Crabs)
> data.plot <- data.frame(width=(18:34))
> lp <- predict(fit, newdata=data.plot, se.fit=TRUE)
> pred.prob <- exp(lp$fit)/(1 + exp(lp$fit))
> LB <- lp$fit - qnorm(0.975)*lp$se.fit
> UB <- lp$fit + qnorm(0.975)*lp$se.fit
> LB.p <- exp(LB)/(1 + exp(LB)); UB.p <- exp(UB)/(1 + exp(UB))
> lines(18:34, pred.prob)
> lines(18:34, LB.p, col="red"); lines(18:34, UB.p, col="blue")
-------------------------------------------------------------------------------
```

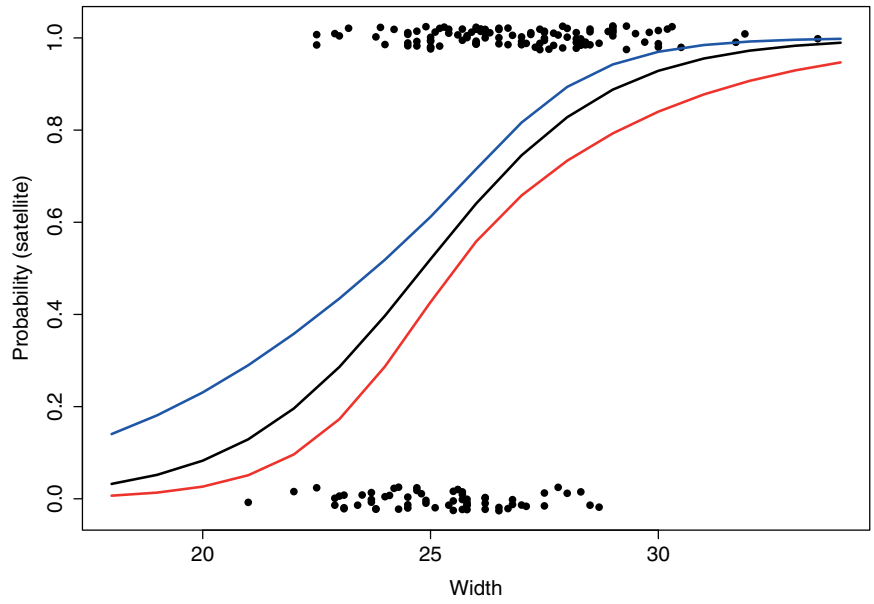


Figure 4.3 Prediction equation and 95% confidence bands for probability of a horseshoe crab satellite as a function of shell width.

4.2.4 Why Use a Model to Estimate Probabilities?

Instead of estimating $P(Y = 1)$ using the logistic model fit, as we just did at $x = 26.5$, we could use the sample proportion to estimate it. Six crabs in the sample had width 26.5, and

four of them had satellites. The sample proportion estimate at $x = 26.5$ is $\hat{\pi} = 4/6 = 0.67$, similar to the model-based estimate.

When the logistic regression model holds, the model-based estimator is much better than the sample proportion. It uses *all* the data rather than only the data at the particular x value. The result is a more precise estimate. For instance, at $x = 26.5$, software reports an $SE = 0.04$ for the model-based estimate 0.695. By contrast, the SE for the sample proportion of 0.67 with only 6 observations is $\sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{(0.67 \times 0.33)/6} = 0.19$. The 95% confidence intervals are (0.61, 0.77) using the model versus (0.30, 0.90) for the score interval using only the 6 observations and sample proportion at $x = 26.5$.

Reality is a bit more complicated. In practice, any model will not *exactly* represent the true relationship between $P(Y = 1)$ and x. If the model approximates the true probabilities reasonably well, however, it performs well. The model-based estimator tends to be much closer than the sample proportion to the true value, unless the sample size on which that sample proportion is based is extremely large. The model smooths the sample data, somewhat dampening the observed variability.

4.3 LOGISTIC REGRESSION WITH CATEGORICAL PREDICTORS

Logistic regression, like ordinary regression, can have multiple explanatory variables. Some or all of them can be categorical, rather than quantitative. This section shows how to include categorical explanatory variables, often called *factors*.

4.3.1 Indicator Variables Represent Categories of Predictors

Suppose that a binary response has two binary explanatory variables, which we denote by x and z. We can then display the data in a $2 \times 2 \times 2$ contingency table, such as analyzed in the next example.

We denote the two categories of x and z by the values 0 and 1. Then x and z are called *indicator variables*,[6] because each indicates the category of an explanatory variable. For the logistic model

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 z$$

with this 0/1 coding, Table 4.1 shows the logit values at the four combinations of values of the two explanatory variables.

Table 4.1 Logits implied by indicator variables in logistic model with two binary factors, $\text{logit}[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 z$.

x	z	Logit
0	0	α
1	0	$\alpha + \beta_1$
0	1	$\alpha + \beta_2$
1	1	$\alpha + \beta_1 + \beta_2$

[6] An alternative name is *dummy variables*.

This model assumes an absence of interaction. The effect of one factor is the same at each category of the other factor. At a fixed category z, the effect on the logit of changing from $x = 0$ to $x = 1$ is

$$= [\alpha + \beta_1(1) + \beta_2 z] - [\alpha + \beta_1(0) + \beta_2 z] = \beta_1.$$

This difference between two logits equals the difference of log odds. Equivalently, that difference equals the log odds ratio between x and y, at that category z. Thus, $\exp(\beta_1)$ equals the conditional odds ratio between x and y. For each category z, the odds of *success* at $x = 1$ equal $\exp(\beta_1)$ times the odds of success at $x = 0$. The lack of an interaction term implies a common value of the odds ratio for the partial tables at the two categories of z. The model satisfies *homogeneous association* (Section 2.7.5).

4.3.2 Example: Survey about Marijuana Use

Table 4.2 is from a survey that asked students in their final year of a high school near Dayton, Ohio, whether they had ever used marijuana. The explanatory variables are gender and race.

Table 4.2 Use of marijuana by gender and race.

		Marijuana Use	
Race	Gender	Yes	No
White	Female	420	620
	Male	483	579
Other	Female	25	55
	Male	32	62

Source: Thanks to Prof. Harry Khamis, Wright State University, for supplying these data, which are in the `Marijuana` data file at the text website.

Following are some results of fitting the logistic model with main effects of gender and race for predicting marijuana use (1 = yes, 0 = no), using the grouped data file:

```
------------------------------------------------------------------------------
> Marijuana <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Marijuana.dat",
+                         header=TRUE)
> Marijuana
   race gender yes  no
1 white female 420 620
2 white   male 483 579
3 other female  25  55
4 other   male  32  62
> fit <- glm(yes/(yes+no) ~ gender + race, weights = yes + no,
+            family=binomial, data=Marijuana)
> summary(fit)
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -0.83035     0.16854   -4.927  8.37e-07
gendermale     0.20261     0.08519    2.378   0.01739
```

```
racewhite       0.44374     0.16766    2.647   0.00813
---
    Null deviance: 12.7528  on 3  degrees of freedom
Residual deviance:  0.0580  on 1  degrees of freedom
-------------------------------------------------------------------------------
```

R has set up indicator variables for x = gender (1 = male, 0 = female) and for z = race (1 = white, 0 = other).

The ML estimated effects are $\hat{\beta}_1 = 0.203$ for gender and $\hat{\beta}_2 = 0.444$ for race. The estimated conditional odds ratio between marijuana use and gender equals $\exp(0.203) = 1.22$. For each race, the estimated odds that a male ($x = 1$) had used marijuana were 1.22 times the estimated odds that a female had used marijuana. For each gender, the estimated odds that a person of white race ($z = 1$) had used marijuana were $\exp(0.444) = 1.56$ times the estimated odds that a person of other races had used marijuana.

The comparison of the null and residual deviances, $12.75 - 0.06 = 12.69$, is the likelihood-ratio statistic with $df = 3 - 1 = 2$ for testing H_0: $\beta_1 = \beta_2 = 0$. This has chi-squared P-value = 0.002. We can conclude that at least one of the explanatory variables has an effect. The hypothesis that gender has no effect on marijuana use, adjusting for race, is H_0: $\beta_1 = 0$. The likelihood-ratio statistic comparing the model with the simpler one having no gender effect equals the difference in residual deviances for the two models. This is 5.67, with $df = 1$, showing evidence of association ($P = 0.017$):

```
-------------------------------------------------------------------------------
> library(car)
> Anova(glm(yes/(yes+no) ~ gender + race, weights=yes+no, family=binomial,
+       data=Marijuana))
       LR Chisq  Df  Pr(>Chisq)
gender   5.6662   1     0.01729  # likelihood-ratio tests for
race     7.2770   1     0.00698  # individual explanatory variables
-------------------------------------------------------------------------------
```

The Wald statistic $z = \hat{\beta}_1/SE = 0.203/0.085 = 2.38$ (or its square of 5.66) provides similar results. The effect of race is also significant.

The small residual deviance of 0.06 with $df = 1$ suggests that the model fits the data adequately. We will address model goodness of fit further in the next chapter (Section 5.2.1).

4.3.3 ANOVA-Type Model Representation of Factors *

A factor having two categories requires only a single indicator variable, taking value 1 or 0 to indicate the category. A factor having c categories requires $c - 1$ indicator variables, as shown below and in Section 4.4.1.

An alternative model representation of factors in logistic regression uses the way ANOVA (analysis of variance) models often express factors. The model formula

$$\text{logit}[P(Y = 1)] = \alpha + \beta_i^x + \beta_k^z$$

represents the effects of x through parameters $\{\beta_i^x\}$ and the effects of z through parameters $\{\beta_k^z\}$. (The x and z superscripts are merely labels and do not represent powers.) The term

β_i^x denotes the effect on Y of classification in category i of x. The variable x has no effect on Y, adjusting for z, when $\beta_1^x = \beta_2^x = \cdots = \beta_c^x$.

This model form applies for any numbers of categories for x and z. Each factor has as many parameters as it has categories, but one is redundant and can be equated to 0. If x has c levels, it has $c - 1$ nonredundant parameters. When we set $\beta_1^x = 0$, for instance, the term β_i^x in this model formula is a simple way of representing

$$\beta_2^x x_2 + \beta_3^x x_3 + \cdots + \beta_c^x x_c,$$

where $(x_2, \ldots, x_c)$ are indicator variables for each category except the first; e.g., $x_2 = 1$ when an observation is in category 2 and $x_2 = 0$ otherwise, and so forth. Category 1 does not need an indicator, because we know an observation is in that category when $x_2 = \cdots = x_c = 0$. This is the type of coding used by R, with its *factor* statement. Some software[7] instead sets $\beta_c^x = 0$. With binary x, the difference $\beta_1^x - \beta_2^x$ is the same for each coding scheme and represents the conditional log odds ratio between x and y, given z. For example, from the output in Section 4.3.2 after Table 4.2 for this example, for gender, $\hat{\beta}_1^x = 0$ and $\hat{\beta}_2^x = 0.203$. Therefore, for each race, the estimated odds that a male (category 2 of gender) had used marijuana were $\exp(\hat{\beta}_2^x - \hat{\beta}_1^x) = \exp(0.203) = 1.22$ times the estimated odds that a female had used marijuana.

By itself, the parameter estimate for a single category of a factor is irrelevant. Different ways of handling parameter redundancies result in different values for that estimate. An estimate makes sense only by comparison with one for another category. Exponentiating a *difference* between estimates for two categories determines the odds ratio relating to the effect of classification in one category rather than the other.

4.3.4 Tests of Conditional Independence and of Homogeneity for Three-Way Contingency Tables*

In many examples with two categorical predictors, x identifies two groups to compare and z is a control variable or a potential confounder. For example, in a clinical trial x might refer to two treatments and z might refer to several centers that recruited patients for the study. Exercise 4.14 shows such an example. The data then can be presented in several 2×2 contingency tables.

With L categories (layers) for z, the model refers to a $2 \times 2 \times L$ contingency table. We can express the model as

$$\text{logit}[P(Y = 1)] = \alpha + \beta x + \beta_k^z,$$

where x is an indicator variable. Then, $\exp(\beta)$ is the common odds ratio between x and y for each of the L partial tables for categories of z. This is the *homogeneous association* structure for multiple 2×2 tables, introduced in Section 2.7.5. In this model, conditional independence between x and y, adjusting for z, corresponds to $\beta = 0$ and an odds ratio of 1 for each partial table. We can test conditional independence by the likelihood-ratio test or the Wald test of H_0: $\beta = 0$. Some software also reports a third test for this hypothesis, called the *Cochran–Mantel–Haenszel test*. This test was proposed in 1959, well before logistic regression was popular, and its formula seems to have nothing to do with modeling. In fact, however, it is the score test (Section 3.4.1) of H_0: $\beta = 0$.

[7] Such as SAS with its *CLASS* statement.

This model has the homogeneous association property. Sometimes it is of interest to test the hypothesis of homogeneous association. A test of homogeneity of the odds ratios is, equivalently, a test of the goodness of fit of the model. The residual deviance does this, as we explain further in Section 5.2.1.

4.4 MULTIPLE LOGISTIC REGRESSION

The general logistic regression model has multiple explanatory variables that can be quantitative, categorical, or both. For p explanatory variables, the model for the log odds is

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

The parameter β_j refers to the effect of x_j on the log odds that $Y = 1$, adjusting for the other x's. For example, $\exp(\beta_1)$ is the multiplicative effect on the odds of a 1-unit increase in x_1, at a fixed value for $\beta_2 x_2 + \cdots + \beta_p x_p$, such as when we can hold constant $x_2, \ldots, x_p$.

4.4.1 Example: Horseshoe Crabs with Color and Width Predictors

We continue the analysis of the horseshoe crab data (Section 4.1.3) by using both the female crab's shell width (quantitative) and color (categorical) as explanatory variables. Color has categories (1 = medium light, 2 = medium, 3 = medium dark, 4 = dark). Color is a surrogate for age, with older crabs tending to have a darker shell color. It is an ordinal variable, but we first treat it as a nominal-scale factor by using three indicator variables for the four color categories. The model is

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4, \tag{4.3}$$

where x denotes the shell width and

$c_2 = 1$ for color = medium, 0 otherwise,
$c_3 = 1$ for color = medium dark, 0 otherwise,
$c_4 = 1$ for color = dark, 0 otherwise.

Here is edited R output for the model fit:

```
------------------------------------------------------------------------------
> fit <- glm(y ~  width + factor(color), family=binomial, data=Crabs)
> summary(fit)
                Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)    -11.38519     2.87346   -3.962  7.43e-05
width            0.46796     0.10554    4.434  9.26e-06
factor(color)2   0.07242     0.73989    0.098     0.922
factor(color)3  -0.22380     0.77708   -0.288     0.773
factor(color)4  -1.32992     0.85252   -1.560     0.119
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.46  on 168  degrees of freedom
------------------------------------------------------------------------------
```

For instance, for medium-light colored crabs (category 1), $c_2 = c_3 = c_4 = 0$, and the prediction equation is $\text{logit}[\hat{P}(Y = 1)] = -11.385 + 0.468x$. By contrast, for dark crabs, $c_2 = c_3 = 0$ and $c_4 = 1$, so

$$\text{logit}[\hat{P}(Y = 1)] = (-11.385 - 1.330) + 0.468x = -12.715 + 0.468x.$$

The model assumes a lack of interaction between width and color. Width has the same effect, with coefficient 0.468, for all colors. This implies that the shapes of the four curves relating width to $P(Y = 1)$ for the four colors are identical. For each color, a 1-cm increase in width has a multiplicative effect of $\exp(0.468) = 1.60$ on the odds that $Y = 1$. Figure 4.4 displays the fitted model. Any one curve is any other curve shifted to the right or to the left.

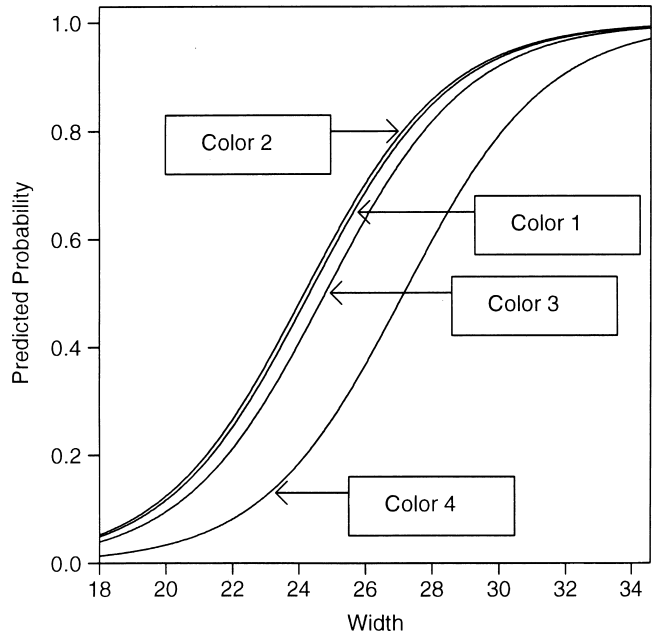


Figure 4.4 Logistic regression model for horseshoe crab satellites using width (quantitative) and color (categorical) explanatory variables.

The parallelism of curves in the horizontal dimension implies that two curves never cross. At all width values, for example, color 4 (dark) has a lower estimated probability of a satellite than the other colors. To illustrate, a dark crab of average width (26.3 cm) has estimated probability

$$\exp[-12.715 + 0.468(26.3)]/\{1 + \exp[-12.715 + 0.468(26.3)]\} = 0.399.$$

By contrast, a medium-light crab of average width has estimated probability

$$\exp[-11.385 + 0.468(26.3)]/\{1 + \exp[-11.385 + 0.468(26.3)]\} = 0.715.$$

The exponentiated difference between two color parameter estimates is an odds ratio comparing those colors. For example, the difference in color parameter estimates between

medium-light crabs and dark crabs equals 1.330. Therefore, at any fixed value of width, the estimated odds that a medium-light crab has a satellite are $\exp(1.330) = 3.8$ times the estimated odds for a dark crab. Using the probabilities just calculated at width 26.3, the odds equal 0.399/0.601 = 0.66 for a dark crab and 0.715/0.285 = 2.51 for a medium-light crab, for which 2.51/0.66 = 3.8.

4.4.2 Model Comparison to Check Whether a Term is Needed

Are certain terms needed in a model? To test this, we can compare the deviance values for that model and for the simpler model without those terms.

For example, to test whether color contributes significantly to model (4.3), we test H_0: $\beta_2 = \beta_3 = \beta_4 = 0$. This hypothesis states that, adjusting for width, the presence of at least one satellite is independent of color. The likelihood-ratio test compares the maximized log-likelihood L_1 for the full model (4.3) to the maximized log-likelihood L_0 for the simpler model in which those parameters equal 0, the fit of which is shown again in the following output. The test statistic $2(L_1 - L_0)$ is identical to the difference between the deviances for the two models, $194.45 - 187.46 = 6.99$. Under H_0, this test statistic has an approximate chi-squared distribution with $df = 3$, the difference between the numbers of parameters in the two models. The P-value of 0.07 provides slight evidence of a color effect. We can obtain the test in R either by comparing the model deviances or by using the `Anova` function in the `car` package:

```
-------------------------------------------------------------------------------
> summary(glm(y ~ width, family=binomial, data=Crabs))
             Estimate  Std. Error  z value   Pr(>|z|)
(Intercept)  -12.3508      2.6287   -4.698   2.62e-06
width          0.4972      0.1017    4.887   1.02e-06
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of freedom # deviance=187.46 when
                                                     # color also in model
> library(car)
> Anova(glm(y ~ width + factor(color), family=binomial, data=Crabs))
              LR Chisq  Df  Pr(>Chisq)
width           24.6038   1   7.041e-07 # LR test of width effect
factor(color)    6.9956   3     0.07204 # LR test of color effect  .
-------------------------------------------------------------------------------
```

Since the analysis in the previous subsection noted that estimated probabilities are quite different for dark-colored crabs, it seems safest to leave color in the model.

4.4.3 Example: Treating Color as Quantitative or Binary

Color has a natural ordering of categories, from medium light to dark. Model (4.3) ignores this ordering, treating color as nominal scale. A simpler model treats color in a quantitative manner. It supposes a linear effect, on the logit scale, for a set of scores assigned to its categories. It is advantageous to treat ordinal explanatory variables in a quantitative manner, when such models fit well. The model is simpler and easier to interpret, and tests of the

effect of the variable are generally more powerful when it has a single parameter rather than several parameters.

To illustrate, we use scores $c = (1, 2, 3, 4)$ for the color categories and fit the model

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 c. \tag{4.4}$$

```
------------------------------------------------------------------------------
> fit2 <- glm(y ~  width + color, family=binomial, data=Crabs)
> summary(fit2) # color treated as quantitative with scores (1, 2, 3, 4)
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -10.0708      2.8068   -3.588  0.000333
width          0.4583      0.1040    4.406  1.05e-05
color         -0.5090      0.2237   -2.276  0.022860
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 189.12  on 170  degrees of freedom

> anova(fit2, fit, test="LRT") # likelihood-ratio test comparing models
Model 1: y ~  width + color
Model 2: y ~  width + factor(color)
  Resid. Df   Resid. Dev  Df  Deviance  Pr(>Chi)
1       170       189.12
2       168       187.46   2    1.6641    0.4351
------------------------------------------------------------------------------
```

The results show strong evidence of an effect for each predictor. At a given width, for every one-category increase in color darkness, the estimated odds of a satellite multiply by $\exp(-0.509) = 0.60$. For example, the estimated odds of a satellite for dark-colored crabs are 60% of those for medium-dark crabs. Since $\exp[3(-0.509)] = 0.22$, the estimated odds of a satellite for dark-colored crabs are 22% of those for medium-light crabs.

A likelihood-ratio test shown in the output compares the fit to the more complex model that has a separate parameter for each color. The difference between the deviances equals $189.12 - 187.46 = 1.66$, based on $df = 2$ ($P = 0.44$). It tests that the color parameters in model (4.3), when plotted against the color scores, follow a linear trend. The simpler model seems to be adequate.

The estimates of the color parameters in the model (4.3) that treats color as a nominal-scale factor are (0, 0.07, -0.22, -1.33). The 0 value for the first category reflects the lack of an indicator variable for that category. Although these values do not depart significantly from a linear trend, the first three are similar compared to the last one. This suggests that another potential color scoring for model (4.4) is $(0, 0, 0, 1)$; that is, $c = 1$ for dark-colored crabs, and $c = 0$ otherwise. Here is output for this model:

```
------------------------------------------------------------------------------
> Crabs$c4 <- ifelse(Crabs$color == 4, 1, 0) # indicator for color cat. 4
> # or could use I(Crabs$color == 4) to directly define indicator var.
> fit3 <- glm(y ~ width + c4, family=binomial, data=Crabs)
> summary(fit3)
```

```
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -11.6790      2.6925   -4.338  1.44e-05
width          0.4782      0.1041    4.592  4.39e-06
c4            -1.3005      0.5259   -2.473    0.0134
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.96  on 170  degrees of freedom

> anova(fit3, fit, test="LRT") # likelihood-ratio test comparing models
Model 1: y ~ width + c4
Model 2: y ~ width + factor(color)
  Resid. Df  Resid. Dev  Df  Deviance  Pr(>Chi)
1       170      187.96
2       168      187.46   2   0.50085    0.7785
------------------------------------------------------------------------
```

At a given width, the estimated odds that a dark crab has a satellite are $\exp(-1.3005) = 0.27$ times the estimated odds for a lighter-colored crab. The likelihood-ratio statistic comparing this model with binary scoring to model (4.3) with color factor suggests that this simpler model also seems adequate.

In summary, the model with color factor, the model with quantitative color scores $(1, 2, 3, 4)$, and the model with binary color scores $(0, 0, 0, 1)$ all suggest that dark crabs are the least likely to have satellites. When the sample size is not very large, typically several models can fit adequately.

4.4.4 Allowing Interaction between Explanatory Variables

The models fitted so far assume a lack of interaction between width and color. Let us check now whether this is sensible. We can allow interaction by adding cross-products of terms for width and color. Each color then has a different-shaped curve relating width to the probability of a satellite, so a comparison of two colors varies according to the value of width.

To illustrate, we consider the model just fitted that has an indicator variable for dark-colored crabs. We next fit the model with an interaction term:

```
------------------------------------------------------------------------
> glm(y ~ width + c4 + width:c4, family=binomial, data=Crabs)
(Intercept)         width            c4     width:c4
   -12.8117        0.5222        6.9578      -0.3217

Null Deviance:      225.76
Residual Deviance:  186.79
------------------------------------------------------------------------
```

The prediction equation is

$$\text{logit}[\hat{P}(Y = 1)] = -12.812 + 0.522x + 6.958c4 - 0.322(x \cdot c4).$$

For dark-colored crabs, $c4 = 1$ and

$$\text{logit}[\hat{P}(Y = 1)] = -5.854 + 0.200x.$$

For lighter-colored crabs, $c4 = 0$ and

$$\text{logit}[\hat{P}(Y = 1)] = -12.812 + 0.522x.$$

The curve for lighter-colored crabs has a stronger effect for x. The curves cross at x such that $-5.854 + 0.200x = -12.812 + 0.522x$, that is, at $x = 21.6$ cm. The sample widths range between 21.0 and 33.5 cm, so the lighter-colored crabs have a higher estimated probability of a satellite over nearly the entire range.

We can compare this model to the simpler model without interaction to analyze whether the fit is significantly better. The likelihood-ratio statistic comparing the model deviances is $187.96 - 186.79 = 1.17$, with $df = 1$. The evidence of interaction is not strong ($P = 0.28$). Although the sample slopes for the width effect are quite different for the two colors, the sample had only 22 crabs of dark color. Therefore, effects involving it have relatively large standard errors. The no-interaction model has the advantage of simpler interpretation.

4.4.5 Effects Depend on Other Explanatory Variables in Model

Suppose we fit an ordinary linear model with an explanatory variable x_1. If we then add to the linear predictor a variable x_2 that is uncorrelated with x_1, the estimated effect of x_1 does not change. In logistic regression, however, the effect of x_1 does change.

When would the effect of x_1 be the same in the logistic models with and without x_2? One case is when x_2 is *conditionally independent* of x_1, given y, rather than marginally uncorrelated. We discuss this further in Section 7.4, in the context of association structure in contingency tables. For now, however, we note that odds ratio effects need not behave the same way that effects behave in ordinary linear models. Partly because of this, to summarize effects and to compare effects for different models or for different groups, it can be more relevant to describe effects on the probability scale than the logit or odds ratio scale. We show ways of doing this in the next section.

4.5 SUMMARIZING EFFECTS IN LOGISTIC REGRESSION

We have interpreted effects in logistic regression using multiplicative effects on the odds, which correspond to odds ratios. However, many practitioners who need to evaluate results of statistical analyses find it easier to understand probabilities than odds ratios.

4.5.1 Probability-Based Interpretations

One way to describe the effect of an explanatory variable x_j sets the other variables at their sample means and finds $\hat{P}(Y = 1)$ at the smallest and largest x_j values. The effect is summarized by reporting those $\hat{P}(Y = 1)$ values or their difference. For a continuous explanatory variable, a caveat for such measures is that their relevance depends on the plausibility of x_j taking extreme values when all other explanatory variables fall at their means.

Also, this summary can be misleading when outliers exist on x_j, in which case it is more sensible to report the estimated probabilities at the upper and lower quartiles of x_j. The effect then describes the change in $\hat{P}(Y = 1)$ values over the middle 50% of the range of x_j values.

We illustrate for the prediction equation for the horseshoe crab data with width and dark-color predictors (Section 4.4.3), $\text{logit}[\hat{P}(Y = 1)] = -11.68 + 0.478x - 1.300c4$. At the mean width, $\hat{P}(Y = 1) = 0.40$ when $c4 = 1$ and $\hat{P}(Y = 1) = 0.71$ when $c4 = 0$. This color effect, differentiating dark-colored crabs from others, seems to be nonnegligible. At the mean for $c4$, $\hat{P}(Y = 1)$ increases from 0.14 to 0.98 between the minimum and maximum width values. It increases from 0.52 to 0.80 between the lower and upper quartiles. These changes reflect a relatively strong width effect. Since $c4$ takes only values 0 and 1, we could instead report such effects separately for each value of $c4$ rather than just at its mean. Here is how to use R to obtain such summary effects:

```
------------------------------------------------------------------------------
> fit3 <- glm(y ~  width + c4, family=binomial, data=Crabs)
> predict(fit3, data.frame(c4=1, width=mean(Crabs$width)), type="response")
0.40063
> predict(fit3, data.frame(c4=0, width=mean(Crabs$width)), type="response")
0.71047
> predict(fit3,data.frame(c4=mean(c4),width=quantile(Crabs$width)), type="resp")
     0%      25%      50%      75%     100%
0.14164  0.51583  0.65412  0.80256  0.98487
------------------------------------------------------------------------------
```

Table 4.3 summarizes effects using estimated probabilities.

Table 4.3 Summary of effects in logistic model with crab width and indicator for dark color as predictors of presence of satellites.

Variable	Estimate	*SE*	Comparison	Change in Estimated Probability
Width (x)	0.478	0.104	(max, min) at $\overline{c4}$	$0.84 = 0.985 - 0.142$
			(UQ, LQ) at $\overline{c4}$	$0.29 = 0.803 - 0.516$
Color ($c4$)	-1.300	0.526	(0, 1) at $\bar{x}$	$0.31 = 0.710 - 0.401$

4.5.2 Marginal Effects and Their Average

For a relatively small change in a quantitative explanatory variable, Section 4.1.2 used the slope of a straight line to approximate the change in the probability. This simpler interpretation applies also with multiple explanatory variables.

Consider a setting of explanatory variables at which $\hat{P}(Y = 1) = \hat{\pi}$. Then, adjusting for the other explanatory variables, a 1-unit increase in x_j corresponds approximately to a $\hat{\beta}_j\hat{\pi}(1 - \hat{\pi})$ change in $\hat{\pi}$. For example, for the horseshoe crab data with explanatory variables $x =$ width and an indicator $c4$ that is 1 for dark crabs and 0 otherwise, the estimated effect of $x =$ width is $\hat{\beta}_1 = 0.478$. When $\hat{\pi} = 0.50$, the approximate effect on $\hat{\pi}$ of a 1-cm increase in x is $0.478(0.50 \times 0.50) = 0.12$. This is considerable, since a 1-cm change in width is less than half its standard deviation (which is 2.1 cm).

This probability rate of change for describing the effect of an explanatory variable depends on the value of $\hat{\pi}$. An overall summary of this effect averages the rate of change at the n sample values of the explanatory variables. Some software refers to this measure as an *average marginal effect*. For a binary explanatory variable, we can average the difference between the estimate of $P(Y = 1)$ for the two categories, called a *discrete change*. We illustrate for the logistic model with width and binary color explanatory variables:

```
------------------------------------------------------------------------------
> fit3 <- glm(y ~  width + c4, family=binomial, data=Crabs)
> library(mfx)
> logitmfx(fit3, atmean=FALSE, data=Crabs) # with atmean=TRUE, finds
Marginal Effects:                          # effect only at the mean
          dF/dx  Std. Err.         z     P>|z|
width   0.08748    0.02447   3.5748  0.00035
c4     -0.26142    0.10569  -2.4735  0.01338
---
dF/dx is for discrete change for the following variables: "c4"
------------------------------------------------------------------------------
```

At the $n = 173$ observed width values, the average rate of change is 0.087 in the estimated probability of a satellite per 1-cm increase in width, adjusting for color. At those width values, the estimated probability of a satellite is 0.261 lower if the crab has dark color than if it has a lighter color.

4.5.3 Standardized Interpretations

With multiple explanatory variables, can we compare magnitudes of $\{\hat{\beta}_j\}$ to compare the effects of the explanatory variables? Comparing estimated effects for binary explanatory variables compares conditional log odds ratios, adjusting for the other explanatory variables in the model. For quantitative explanatory variables, this is relevant if those variables have the same units, so that a 1-unit change means the same thing for each. Otherwise, we can use *standardized* coefficients. We can obtain these by fitting the model to standardized explanatory variables, replacing each x_j by $(x_j - \bar{x}_j)/s_{x_j}$, where s_{x_j} denotes the sample standard deviation of x_j. A one-unit change in the standardized variable is a standard deviation change in the original variable. Then, $\hat{\beta}_j$ represents the effect of a standard deviation change in x_j, adjusting for the other variables. The standardized effect estimate for x_j is the unstandardized estimate multiplied by s_{x_j}.

To illustrate, the prediction equation in Section 4.4.3 for the horseshoe crab data using width and quantitative color (scores 1, 2, 3, 4) as explanatory variables is $\text{logit}(\hat{\pi}) = -10.071 + 0.458x - 0.509c$. The explanatory variables have quite different variability, so it is not correct to conclude from this equation that the effects have similar magnitudes. Width has $\bar{x} = 26.30$ and $s_x = 2.11$, while color has $\bar{c} = 2.44$ and $s_c = 0.80$. For standardized explanatory variables, the estimated effects equal $(2.11)(0.458) = 0.97$ and $(0.80)(-0.509) = -0.41$. A standard deviation increase in width is estimated to have more than double the effect of a standard deviation increase in color, adjusting for the other variable.

4.6 SUMMARIZING PREDICTIVE POWER: CLASSIFICATION TABLES, ROC CURVES, AND MULTIPLE CORRELATION

For comparing models, it is often useful to summarize their predictive power, that is, how well we can predict the response variable outcome using the model fit. This section shows three ways to do this.

4.6.1 Summarizing Predictive Power: Classification Tables

A *classification table* cross-classifies the binary outcome y with a prediction of whether $y = 0$ or 1. The prediction for observation i is $\hat{y} = 1$ when its estimated probability $\hat{\pi}_i > \pi_0$ and $\hat{y} = 0$ when $\hat{\pi}_i \leq \pi_0$, for some cutoff π_0. One possibility is to take $\pi_0 = 0.50$. However, if a low (high) proportion of observations have $y = 1$, the model fit may never (always) have $\hat{\pi}_i > 0.50$, in which case one never (always) predicts $\hat{y} = 1$. Another possibility takes π_0 as the sample proportion of 1 outcomes, which is $\hat{\pi}_i$ for the model containing only an intercept term.

We illustrate for the model for the presence of horseshoe crab satellites using width and the color factor as explanatory variables (Section 4.4.1). Of the 173 crabs, 111 had a satellite, for a sample proportion of 0.6416:

```
-------------------------------------------------------------------------------
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
+                     header=TRUE)
> prop <- sum(Crabs$y)/nrow(Crabs) # sample proportion of 1's for y variable
> prop
[1] 0.6416185
> fit <- glm(y ~  width + factor(color), family=binomial, data=Crabs)
> predicted <- as.numeric(fitted(fit) > prop) # predict y=1 when est.> 0.6416
> xtabs(~ Crabs$y + predicted)
        predicted  # Classification table with sample proportion cutoff
Crabs$y  0  1
      0 43 19
      1 36 75
-------------------------------------------------------------------------------
```

Of the 62 cases with $y = 0$, the model predicts $\hat{y} = 0$ for 43; of the 111 cases with $y = 1$, the model predicts $\hat{y} = 1$ for 75. Table 4.4 shows classification tables for $\pi_0 = 0.6416$ and for $\pi_0 = 0.50$.

Table 4.4 Classification tables for horseshoe crab data with width and factor color predictors.

	Prediction, $\pi_0 = 0.6416$		Prediction, $\pi_0 = 0.50$		
Actual	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	Total
$y = 1$	75	36	96	15	111
$y = 0$	19	43	31	31	62

Two useful summaries of predictive power are

$$\text{Sensitivity} = P(\hat{y} = 1 \mid y = 1), \quad \text{Specificity} = P(\hat{y} = 0 \mid y = 0).$$

Section 2.1.2 introduced these measures for predictions with diagnostic medical tests. When $\pi_0 = 0.6416$, from Table 4.4 the estimated sensitivity = 75/111 = 0.676 and estimated specificity = 43/62 = 0.694. The overall proportion of correct classifications is (75 + 43)/173 = 0.682. This estimates

$$\begin{aligned} P(\text{correct classif.}) &= P(y = 1 \text{ and } \hat{y} = 1) + P(y = 0 \text{ and } \hat{y} = 0) \\ &= P(\hat{y} = 1 \mid y = 1)P(y = 1) + P(\hat{y} = 0 \mid y = 0)P(y = 0) \\ &= \text{sensitivity}[P(y = 1)] + \text{specificity}[1 - P(y = 1)], \end{aligned}$$

which is a weighted average of sensitivity and specificity.

These sample summaries of predictive power are overly optimistic because they use $\hat{\pi}_i$ from the model fitted to the data set of which y_i was one element. It is better to make the prediction with the *leave-one-out cross-validation* approach by which $\hat{\pi}_i$ is based on the model fitted to the other $n - 1$ observations. When we do this for this model, we obtain a proportion of correct classifications of 0.671. The proportion correct is 0.642 with color alone as a predictor, 0.659 with width alone, and 0.682 with width and an indicator for whether a crab has dark color.

A classification table has limitations: it collapses continuous predictive values $\hat{\pi}$ into binary ones. The choice of π_0 is arbitrary. Results are sensitive to the relative numbers of times that $y = 1$ and $y = 0$.

4.6.2 Summarizing Predictive Power: ROC Curves

A *receiver operating characteristic* (ROC) curve is a plot that shows the sensitivity and the specificity of the predictions for all the possible cutoffs π_0. This curve is more informative than a classification table, because it summarizes predictive power for all possible π_0.

The ROC curve plots sensitivity on the vertical axis versus (1 − specificity) on the horizontal axis. When π_0 gets near 0, almost all predictions are $\hat{y} = 1$; then, sensitivity is near 1, specificity is near 0, and the point for (1 − specificity, sensitivity) has coordinates near (1, 1). When π_0 gets near 1, almost all predictions are $\hat{y} = 0$; then, sensitivity is near 0, specificity is near 1, and the point for (1 − specificity, sensitivity) has coordinates near (0, 0). The ROC curve usually has a concave or nearly concave shape connecting the points (0, 0) and (1, 1).

Figure 4.5 shows the ROC curve for the model for the horseshoe crabs using width and the color factor as predictors. When $\pi_0 = 0.6416$, for example, the specificity = 0.69, the sensitivity = 0.68, and the point plotted for the ROC curve has coordinates (0.31, 0.68).

For a particular value of specificity, better predictive power corresponds to higher sensitivity. Therefore, the better the predictive power, the higher is the ROC curve. Because of this, the area under the curve provides a single value that summarizes predictive power. The greater the area, the better the predictive power. This measure of predictive power is called the *concordance index*. Consider all pairs of observations (i, j) such that $y_i = 1$ and $y_j = 0$. The concordance index estimates the probability that the predictions and the outcomes are *concordant*, which means that the observation with the larger y also has the larger $\hat{\pi}$. A concordance value of 0.50 means predictions were no better than random guessing.

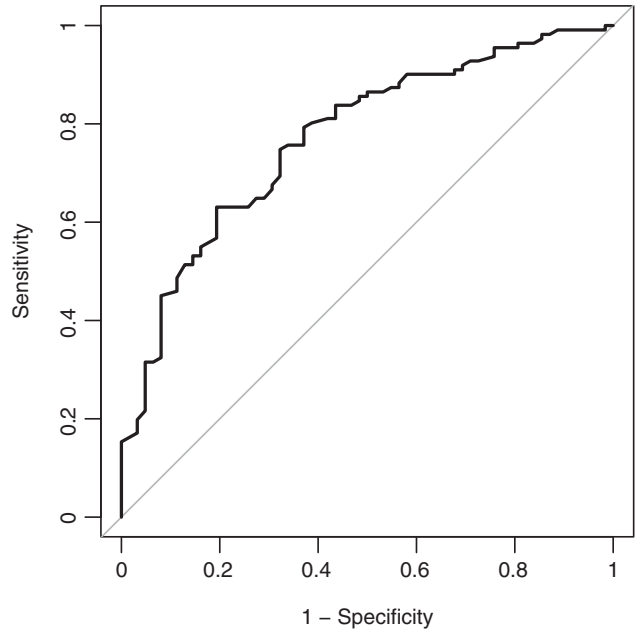


Figure 4.5 ROC curve for logistic regression model with horseshoe crab data and width and color factor predictors.

This corresponds to a model having only an intercept term. Its ROC curve is a straight line connecting the points (0, 0) and (1, 1).

For the horseshoe crab data, the sample concordance index is 0.639 with color alone as a factor predictor, 0.742 with width alone, 0.771 with width and color as a factor, and 0.772 with width and an indicator for whether a crab has dark color. Here is R code for an ROC curve and the area under it:

```
-------------------------------------------------------------------------------
> fit <- glm(y ~ width + factor(color), family=binomial, data=Crabs)
> library(pROC)
> rocplot <- roc(y ~ fitted(fit), data=Crabs)
> plot.roc(rocplot, legacy.axes=TRUE) # Specficity on x axis if legacy.axes=F
> auc(rocplot) # auc = area under ROC curve = concordance index
Area under the curve: 0.7714
-------------------------------------------------------------------------------
```

4.6.3 Summarizing Predictive Power: Multiple Correlation

For a GLM, one summary of prediction power is the correlation R between the observed responses $\{y_i\}$ and the model's fitted values $\{\hat{\mu}_i\}$. For least squares fitting of an ordinary linear model, R is the *multiple correlation* between the response variable and the explanatory variables. Then, R^2 describes the proportion of the variation in y that is explained by

those predictors. An advantage of R compared to R^2 is that it uses the original scale and it has value approximately proportional to the effect size; for instance, with a single quantitative explanatory variable, the correlation is the slope multiplied by the ratio of standard deviations of the two variables.

For a binary regression model, R is the correlation between the n binary $\{y_i\}$ observations (1 or 0 for each) and the fitted probabilities $\{\hat{\pi}_i\}$. The highly discrete nature of y can suppress the range of possible R values, more so when the imbalance between the frequencies of 0 and 1 values is greater. Also, like any correlation measure, the value of R depends on the range of values observed for the explanatory variables. Nevertheless, R is useful for comparing fits of different models for the same data.

Although this measure is not routinely provided by GLM software, it is simple to obtain, as shown here in R code:

```
------------------------------------------------------------------------------
> fit <- glm(y ~ width + factor(color), family=binomial, data=Crabs)
> cor(Crabs$y, fitted(fit))
[1] 0.45221
------------------------------------------------------------------------------
```

The simpler model that uses width and a dark-color indicator does essentially as well, with $R = 0.447$. Using width alone has $R = 0.402$.

The square of this measure does not have the proportional reduction in variation interpretation that it has for ordinary (least squares) regression. Various measures have been proposed in an attempt to do this for binary data. For example, one such measure approximates R^2 for an ordinary regression model presented in Section 5.5.3 for an underlying continuous variable for y. We present this approach for ordinal responses in Section 6.3.7, but it applies also for binary responses.

EXERCISES

4.1 A study[8] investigated characteristics associated with $y =$ whether a cancer patient achieved remission ($1 =$ yes, $0 =$ no). An important explanatory variable was a labeling index ($LI =$ percentage of "labeled" cells) that measures proliferative activity of cells after a patient receives an injection of tritiated thymidine. Table 4.5 shows the data and R output for a logistic regression model.

a. Show that $\hat{P}(Y = 1) = 0.50$ when $LI = 26.0$.

b. When LI increases by 1, show that the estimated odds of remission multiply by 1.16.

c. Summarize the LI effect by how $\hat{P}(Y = 1)$ changes over the range or interquartile range of LI values.

d. Show that the rate of change in $\hat{P}(Y = 1)$ is 0.009 when $LI = 8$.

e. Summarize the LI effect by the estimated average marginal effect.

[8] Article by E.T. Lee, *Computer Prog. Biomed.* **4**: 80–92 (1974).

Table 4.5 Software output for Exercise 4.1 on cancer remission.

```
------------------------------------------------------------------------------
> LI <- c(8,8,10,10,12,12,12,14,14,14,16,16,16,18,20,20,20,22,22,24,26,28,32,34,
+         38,38,38)
> y <- c(0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,1,0,0,1,1,0,1,1,1,0)
> summary(glm(y ~ LI, family=binomial))
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -3.77714     1.37862   -2.740   0.00615
LI            0.14486     0.05934    2.441   0.01464
---
    Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 26.073  on 25  degrees of freedom

>  confint(glm(y ~ LI, family=binomial))
                2.5 %     97.5 %
LI            0.04252    0.28467
------------------------------------------------------------------------------
```

4.2 Refer to the previous exercise. Using information from Table 4.5:

a. Conduct a Wald test for the LI effect and construct a 95% Wald confidence interval for the odds ratio corresponding to a 1-unit increase in LI. Interpret.

b. Conduct a likelihood-ratio test and construct a 95% profile likelihood interval. Interpret.

4.3 Refer to the previous two exercises. Set up the data file as 14 observations in grouped-data format. Compare to the `Remissions` data file at the text website. Fit the model with this data file. Are the ML model parameter estimates the same as with the ungrouped data file? Is the deviance the same? Why or why not?

4.4 For the snoring and heart disease data of Table 3.1 (Section 3.2.3) with snoring-level scores (0, 2, 4, 5), the logistic regression ML fit is $\text{logit}[\hat{P}(Y = 1)] = -3.866 + 0.397x$. Interpret the effect of snoring on the odds of heart disease.

4.5 For the 23 space shuttle flights before the Challenger mission disaster in 1986, Table 4.6 and the `Shuttle` data file at the text website shows the temperature (°F)

Table 4.6 Data for Exercise 4.5 on space shuttle.

Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD
1	66	0	2	70	1	3	69	0	4	68	0	5	67	0
6	72	0	7	73	0	8	70	0	9	57	1	10	63	1
11	70	1	12	78	0	13	67	0	14	53	1	15	67	0
16	75	0	17	70	0	18	81	0	19	76	0	20	79	0
21	75	1	22	76	0	23	58	1						

Note: Ft = flight no., Temp = temperature, TD = thermal distress (1 = yes, 0 = no).
Source: Data based on Table 1 in *J. Amer. Statist. Assoc*, **84**: 945–957 (1989), by S.R. Dalal, E.B. Fowlkes, and B. Hoadley. Reprinted with permission from the *J. Amer. Statist. Assoc.*

at the time of the flight and whether at least one primary O-ring suffered thermal distress.

a. Use logistic regression to model the effect of temperature on the probability of thermal distress. Interpret the effect.

b. Estimate the probability of thermal distress at 31°, the temperature at the time of the Challenger flight.

c. At what temperature does the estimated probability equal 0.50? At that temperature, give a linear approximation for the change in the estimated probability per degree increase in temperature.

d. Interpret the effect of temperature on the odds of thermal distress.

e. Test the hypothesis that temperature has no effect.

4.6 For Exercise 3.9 on travel credit cards, use the logistic output there to (**a**) interpret the effect of income on the odds of possessing a travel credit card, and conduct (**b**) a significance test and (**c**) a confidence interval for that effect.

4.7 Hastie and Tibshirani (1990, p. 282) described a study to determine risk factors for kyphosis, which is severe forward flexion of the spine following corrective spinal surgery. The `Kyphosis` data file at the text website shows the 40 observations on $y =$ whether kyphosis is present (1 = yes), with $x =$ age as the explanatory variable.

a. Fit a logistic regression model. Test the effect of age.

b. Plot the data. Note the difference in dispersion of age at the two levels of kyphosis.

c. Fit the model $\text{logit}[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 x^2$. Test the significance of the squared age term, plot the fit, and interpret. (The final paragraph of Section 4.1.5 is relevant to these results.)

4.8 For the `Crabs` data file at the text website, fit the logistic regression model for the probability of a satellite ($y = 1$) using $x =$ weight as the sole explanatory variable.

a. Report the ML prediction equation. At the mean weight value of 2.437 kg, give a linear approximation for the estimated effect of (i) a 1-kg increase in weight. This represents a relatively large increase, so convert this to the effect of (ii) a 0.10-kg increase, (iii) a standard deviation increase in weight (0.58 kg).

b. Find and interpret the average marginal effect of weight per 0.10-kg increase.

c. Construct the classification table using the sample proportion of $y = 1$ as the cutoff. Report the sensitivity and specificity. Interpret.

d. Construct an ROC curve, and report and interpret the area under it.

4.9 For the `Crabs` data file, fit a logistic regression model for the probability of a satellite, using color alone as the predictor.

a. Treat color as a nominal-scale factor. Report the prediction equation and explain how to use it to compare the first and fourth colors.

b. For the model in (**a**), conduct a likelihood-ratio test of the hypothesis that color has no effect. Interpret.

c. Treating color in a quantitative manner (scores 1, 2, 3, 4), obtain a prediction equation. Interpret the coefficient of color and test the hypothesis that color has no effect.

d. When we treat color as quantitative instead of qualitative, state a potential advantage relating to power and a potential disadvantage relating to model lack of fit.

e. Using weight and quantitative color as explanatory variables, find standardized coefficients, and interpret.

4.10 In a study[9] on the effects of AZT in slowing the development of AIDS symptoms, 338 veterans whose immune systems were beginning to falter after infection with the AIDS virus were randomly assigned either to receive AZT immediately or to wait until their T cells showed severe immune weakness. Output follows of modeling the $2 \times 2 \times 2$ cross-classification of race, whether AZT was given immediately, and whether AIDS symptoms developed during the three-year study.

```
------------------------------------------------------------------------------
> AIDS # Data file at text website
   race azt yes no # yes and no are categories of AIDS symptoms response
1 white yes  14 93
2 white  no  32 81
3 black yes  11 52
4 black  no  12 43
> fit <- glm(yes/(yes+no) ~ azt + race, weights=yes+no, family=binomial,
+            data=AIDS)
> summary(fit)
                  Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)       -1.07357     0.26294   -4.083  4.45e-05
aztyes            -0.71946     0.27898   -2.579   0.00991
racewhite          0.05548     0.28861    0.192   0.84755
---
    Null deviance: 8.3499  on 3  degrees of freedom
Residual deviance: 1.3835  on 1  degrees of freedom
------------------------------------------------------------------------------
```

a. What null hypothesis is tested by the difference between the null deviance and the residual deviance? Interpret.

b. Explain how to set up indicator variables for azt and race to obtain the estimates shown.

4.11 For Table 2.9 on racial characteristics and the death penalty, create a data file and fit a logistic model for death penalty as the response, with defendant's race and victims' race as predictors.

a. Report the model fit and interpret the parameter estimates. Based on those estimates, which group is most likely to have the *yes* response?

b. Conduct inference about the effect of victims' race, controlling for defendant's race. Interpret.

4.12 At the website `www.stat.ufl.edu/~aa/intro-cda/data` for the 2nd edition of this book, the `MBTI` data file cross-classifies a sample of people from the MBTI Step II National Sample on whether they report drinking alcohol frequently and on the four

[9] Described in the *New York Times*, Feb. 15, 1991.

binary scales of the Myers–Briggs personality test: Extroversion/Introversion (E/I), Sensing/iNtuitive (S/N), Thinking/Feeling (T/F) and Judging/Perceiving (J/P). The 16 predictor combinations correspond to the 16 personality types: ESTJ, ESTP, ESFJ, ESFP, ENTJ, ENTP, ENFJ, ENFP, ISTJ, ISTP, ISFJ, ISFP, INTJ, INTP, INFJ, INFP. (e.g., of the 77 people of type ESTJ, 13 reported smoking frequently.) Fit a model using the four scales as predictors of the probability of drinking alcohol frequently. Report the prediction equation, specifying how you set up the indicator variables. Based on the model parameter estimates, explain why the personality type with the highest estimated probability of drinking alcohol is ENTP.

4.13 For first-degree murder convictions[10] in East Baton Rouge Parish, Louisiana, between 1990 and 2008, the death penalty was given in 3 out of 25 cases in which a white killed a white, in 0 out of 3 cases in which a white killed a black, in 9 out of 30 cases in which a black killed a white, and in 11 out of 132 cases in which a black killed a black. Table 4.7 shows software output for fitting a logistic regression model, where $d = 1$ ($d = 0$) for black (white) defendants and $v = 1$ ($v = 0$) for black (white) victims. Summarize in a short, nontechnical report what you learn from this output.

Table 4.7 Logistic regression fit of death penalty data for Exercise 4.13.

```
-----------------------------------------------------------------------
             Estimate   Std. Error   z value   Pr(>|z|)
(Intercept)   -2.0232       0.6137    -3.297   0.000978
d              1.1886       0.7236     1.643   0.100461
v             -1.5713       0.5028    -3.125   0.001778
Residual deviance: 0.16676  on 1  degrees of freedom
-----------------------------------------------------------------------
```

4.14 Table 4.8 shows results of an eight-center clinical trial to compare a drug to placebo for curing an infection. At each center, subjects were randomly assigned to treatments.

Table 4.8 Clinical trial data for Exercise 4.14.

		Response				Response	
Center	Treatment	Success	Failure	Center	Treatment	Success	Failure
1	Drug	11	25	5	Drug	6	11
	control	10	27		control	0	12
2	Drug	16	4	6	Drug	1	10
	control	22	10		control	0	10
3	Drug	14	5	7	Drug	1	4
	control	7	12		control	1	8
4	Drug	2	14	8	Drug	4	2
	control	1	16		control	6	1

Source: P.J. Beitler and J.R. Landis, *Biometrics*, vol. 41, pp. 991–1000 (1985).

[10] From G. Pierce and M. Radelet, *Louisiana Law Review*, vol. 71 (2011), pp. 647–673.

Analyze these data, available in the `Infection` data file at the text website. Using logistic regression, describe and make inference about the treatment effect.

4.15 In a study designed to evaluate whether an educational program makes sexually active adolescents more likely to obtain condoms, adolescents were randomly assigned to two experimental groups. The educational program, involving a lecture and videotape about transmission of the HIV virus, was provided to one group but not the other. In logistic regression models, factors observed to influence a teenager to obtain condoms were gender, socioeconomic status (SES), lifetime number of partners, and the experimental group. Table 4.9 summarizes study results.

a. Find the parameter estimates for the fitted model, using (1, 0) indicator variables for the first three explanatory variables.

b. Explain why either the estimate of 1.38 for the odds ratio for gender or the corresponding confidence interval is incorrect. Show that if the reported interval is correct, then 1.38 is actually the *log* odds ratio, and the estimated odds ratio equals 3.98.

Table 4.9 Table for Exercise 4.15 on condom use.

Variables	Odds Ratio	95% Confidence Interval
Group (Education vs. None)	4.04	(1.17, 13.9)
Gender (Males vs. Females)	1.38	(1.23, 12.88)
SES (High vs. Low)	5.82	(1.87, 18.28)
Lifetime no. of Partners	3.22	(1.08, 11.31)

Source: Rickert *et al., Clin. Pediatrics*, 205–210 (1992).

4.16 Table 4.10, which is the `SoreThroat` data file at the text website, shows results of a study about $y =$ whether a patient having surgery with general anesthesia experienced a sore throat on waking (1 = yes, 0 = no) as a function of $d =$ duration of the surgery

Table 4.10 Data for Exercise 4.16 on sore throat after surgery.

Patient	d	t	y	Patient	d	t	y	Patient	d	t	y
1	45	0	0	2	15	0	0	3	40	0	1
4	83	1	1	5	90	1	1	6	25	1	1
7	35	0	1	8	65	0	1	9	95	0	1
10	35	0	1	11	75	0	1	12	45	1	1
13	50	1	0	14	75	1	1	15	30	0	0
16	25	0	1	17	20	1	0	18	60	1	1
19	70	1	1	20	30	0	1	21	60	0	1
22	61	0	0	23	65	0	1	24	15	1	0
25	20	1	0	26	45	0	1	27	15	1	0
28	25	0	1	29	15	1	0	30	30	0	1
31	40	0	1	32	15	1	0	33	135	1	1
34	20	1	0	35	40	1	0				

Source: Data from D. Collett, pp. 350–358 in *Encyclopedia of Biostatistics* (Wiley: 1998). Predictors are $d =$ duration of surgery, $t =$ type of device.

(in minutes) and t = type of device used to secure the airway (0 = laryngeal mask airway, 1 = tracheal tube).

a. Fit a model permitting interaction between the explanatory variables. Report and interpret the prediction equation for the effect of d when (i) $t = 1$, (ii) $t = 0$. Conduct inference about whether you need the interaction term.

b. Compare the predictive power of models with and without the interaction term by finding the correlation R between the observed and fitted values for each model.

4.17 Table 4.11 shows estimated effects for a logistic regression model for y = presence of squamous cell esophageal cancer (1 = yes, 0 = no). Smoking status (s) equals 1 for at least one pack per day and 0 otherwise, alcohol consumption (a) equals the average number of alcoholic drinks consumed per day, and race (r) equals 1 for blacks and 0 for whites.

a. To describe the race-by-smoking interaction, construct the prediction equation when $r = 1$ and again when $r = 0$. Find the fitted conditional odds ratio for the smoking effect for each case. Similarly, construct the prediction equation when $s = 1$ and again when $s = 0$. Find the fitted conditional odds ratio for the race effect for each case. (For each association, the coefficient of the cross-product term is the difference between the log odds ratios at the two levels for the other variable.)

b. In Table 4.11, what do the coefficients of smoking and race represent? What hypotheses do their P-values refer to?

Table 4.11 Table for Exercise 4.17 on effects on esophageal cancer.

Variable	Effect	P-value
Intercept	−7.00	<0.01
Alcohol use	0.10	0.03
Smoking	1.20	<0.01
Race	0.30	0.02
Race × Smoking	0.20	0.04

4.18 For Table 4.12 from the 2016 General Social Survey, create a data file and analyze the data using logistic regression. Summarize your analyses in a short report, including edited output in an appendix.

Table 4.12 Data on belief in afterlife for Exercise 4.18.

		Belief in Afterlife	
Race	Religion	Yes	No or Undecided
White	Protestant	817	250
	Catholic	519	194
	Other	48	9
Black	Protestant	298	86
	Catholic	39	13
	Other	119	38

4.19 For model (4.3) for the horseshoe crabs with color and width predictors, add three terms to permit interaction between color and width.

a. Report the prediction equations relating width to the probability of a satellite, for each color. Plot or sketch them, and interpret.

b. Test whether the interaction model fits better than the simpler model without interaction terms. Interpret. Compare their predictive power by finding the correlation R between the observed and fitted values for each model.

4.20 Refer to Exercise 4.12 about MBTI and alcohol drinking.

a. When the sample proportion of 0.092 who reported drinking alcohol frequently is the cutpoint for forming a classification table, sensitivity $= 0.53$ and specificity $=$ 0.66. Explain what these mean, and show that the sample proportion of correct classifications was 0.65.

b. The `MBTI` data file also shows responses on whether a person smokes frequently. When a classification table for the model containing the four main effect terms to predict smoking uses the sample proportion of frequent smokers of 0.23 as the cutoff, sensitivity $= 0.48$ and specificity $= 0.55$. The area under the ROC curve is 0.55. Does knowledge of personality type help you predict well whether someone is a frequent smoker? Explain.

4.21 Explain how the classification table in Table 4.4 with $\pi_0 = 0.50$ was constructed. Estimate the sensitivity and specificity, and interpret.

4.22 You plan to study the relation between $x =$ age and $y =$ whether a member of Facebook ($1 =$ yes, $0 =$ no). A priori, you predict that $P(Y = 1)$ is currently about 0.80 at $x = 18$ and about 0.20 at $x = 65$. Assuming that the logistic regression model describes this relation well, approximate the value for the effect β of x in the model.

4.23 Table 7.8 in Chapter 7 shows data from the `Substance2` data file at the text website. Create a new data file from which you can use logistic regression to analyze these data, treating marijuana use as the response variable and alcohol use, cigarette use, gender, and race as explanatory variables. Prepare a short report summarizing model-based descriptive and inferential results.

4.24 Refer to the following artificial data:

x	number of trials	number of successes
0	4	1
1	4	2
2	4	4

Denote by M_0 the logistic null model and by M_1 the model that also has x as a predictor. Denote the maximized log-likelihood values by L_0 for M_0, L_1 for M_1, and L_s for the saturated model. Create a data file in two ways, entering the data as (i) ungrouped data: 12 individual binary observations, (ii) grouped data: 3 summary binomial observations each with sample size $= 4$.

a. Fit M_0 and M_1 for each data file. Report L_0 and L_1 (or $-2L_0$ and $-2L_1$) in each case. Do they depend on the form of data entry?

b. Show that the deviances for M_0 and M_1 depend on the form of data entry. Why is this? (*Hint*: The saturated model has 12 parameters for data file (i) but 3 parameters for data file (ii).)

c. Show that the difference between the deviances does not depend on the form of data file. Thus, for testing the effect of x, it does not matter how you enter the data.

CHAPTER 5

BUILDING AND APPLYING LOGISTIC REGRESSION MODELS

Having introduced logistic regression, we now consider how to build a logistic model and check its fit, when possibly many explanatory variables are available as potential predictors. As in ordinary regression, a variety of strategies are possible for selecting the explanatory variables, including stepwise algorithms. After choosing a preliminary model, model checking explores lack of fit, such as goodness-of-fit tests and residuals. Some parameter estimates can be infinite with small or highly unbalanced samples. We will see how this can happen and present possible remedies. We also present alternative link functions such as the probit and alternative methods of inference such as the Bayesian approach.

5.1 STRATEGIES IN MODEL SELECTION

For a data set with a binary response variable and many potential explanatory variables, how do we build a logistic regression model? As in ordinary regression modeling, selecting explanatory variables becomes more challenging as their number increases, because of the rapid increase in possible effects and interactions. The model should be complex enough to fit the data well, but simpler models are easier to interpret.

Most studies are designed to answer certain questions that motivate including certain terms in the model. To answer those questions, *confirmatory* analyses use a restricted set of models. A study's theory about an effect may be tested by comparing models with and without that effect. Whenever one model is a special case of another, we can test that the simpler model is adequate against the alternative that at least one of the extra parameters in the

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

more complex model is nonzero. As Section 3.4.4 showed, the likelihood-ratio test statistic for comparing the models is the difference between the deviances for the models and has an approximate chi-squared distribution. In the absence of underlying theory, some studies are *exploratory* rather than confirmatory. Then, a search among many models may provide clues about which explanatory variables have substantive association with the response and may suggest effects to investigate in future research studies.

5.1.1 How Many Explanatory Variables Can the Model Handle?

Data are unbalanced on the response variable if $y = 1$ or $y = 0$ relatively few times. This limits the number of explanatory variables p for which effects can be estimated precisely. One guideline[1] suggests that the data set should contain at least 10 outcomes of each type for every explanatory variable. For example, if $y = 1$ only 30 times out of $n = 1000$ observations, the model should have no more than $p = 3$ explanatory variables even though the overall sample size is quite large. This guideline is simplistic and a bit conservative. When not satisfied, software still can usually fit the model. In practice, often p is quite large, sometimes even of similar magnitude as the number of observations n. However, when this guideline is violated badly, ML estimators of effects and standard errors may be highly biased.

Cautions that apply to building ordinary regression models hold for any GLM. For example, models can suffer from *multicollinearity* — correlations among explanatory variables making it seem that no one variable is important when all the others are in the model. A variable may seem to have little effect because it overlaps considerably with others in the model, itself being predicted well by the other explanatory variables. Deleting such a redundant variable can be helpful, for instance to reduce standard errors of other estimated effects.

5.1.2 Example: Horseshoe Crab Satellites Revisited

We analyzed the horseshoe crabs data set (Table 3.2) throughout Chapter 4 in terms of whether each of $n = 173$ female crabs has at least one male satellite that could mate with her ($y = 1$) or has no satellites ($y = 0$). The `Crabs` data file has four explanatory variables: color (four categories), spine condition (three categories), shell width, and weight. We now explore logistic regression modeling using all of them.

We first fit a model that contains all the main effects, treating color and spine condition as qualitative (nominal-scale) factors. Let (c_2, c_3, c_4) be indicator variables for 3 of the 4 colors and let (s_2, s_3) be indicator variables for 2 of the 3 spine conditions, suppressing the indicator for the first category (as `R` does). The model is

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 \text{weight} + \beta_2 \text{width} + \beta_3 c_2 + \beta_4 c_3 + \beta_5 c_4 + \beta_6 s_2 + \beta_7 s_3.$$

Here are some results, using `R`:

```
------------------------------------------------------------------------------
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
+                     header=TRUE)
> fit <- glm(y ~ weight + width + factor(color) + factor(spine),
+            family=binomial, data=Crabs)
```

[1] See P. Peduzzi et al., *J. Clin. Epidem.* **49**: 1373–1379 (1996).

```
> summary(fit)
                Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)     -8.06501     3.92855   -2.053    0.0401
weight           0.82578     0.70383    1.173    0.2407
width            0.26313     0.19530    1.347    0.1779
factor(color)2  -0.10290     0.78259   -0.131    0.8954
factor(color)3  -0.48886     0.85312   -0.573    0.5666
factor(color)4  -1.60867     0.93553   -1.720    0.0855 .
factor(spine)2  -0.09598     0.70337   -0.136    0.8915
factor(spine)3   0.40029     0.50270    0.796    0.4259
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 185.20  on 165  degrees of freedom
AIC: 201.2
> 1 - pchisq(225.76-185.20, 172-165) # P-value for test that all beta's = 0
[1] 9.83292e-07
> library(car)
> Anova(fit) # likelihood-ratio tests for individual explanatory variables
              LR Chisq  Df  Pr(>Chisq)
weight          1.4099   1     0.23507
width           1.7968   1     0.18010
factor(color)   7.5958   3     0.05515 .
factor(spine)   1.0091   2     0.60377
-------------------------------------------------------------------------
```

A likelihood-ratio test that Y is jointly independent of the four explanatory variables simultaneously tests H_0: $\beta_1 = \cdots = \beta_7 = 0$. The test statistic is the difference between the null deviance and the residual deviance, which is $225.76 - 185.20 = 40.56$ with $df = 172 - 165 = 7$. This shows extremely strong evidence that at least one explanatory variable has an effect ($P < 0.0001$).

Although this overall test is highly significant, the test results for individual variables are discouraging. The estimates for weight and width are only slightly larger than their *SE* values. The estimates for the factors compare each category to the first one as a baseline. For color, the largest difference is less than two standard errors. For the spine condition, the largest difference is less than a standard error. Likelihood-ratio tests for individual explanatory variables show nothing significant at the 0.05 level (although color comes close). The P-value for the overall test is small, yet the lack of significance for individual effects is a warning sign of multicollinearity. Section 4.2.2 showed strong evidence of a width effect. Adjusting for weight, color, and spine condition, only very weak evidence remains of a width effect. However, weight and width have a strong correlation (0.887). For practical purposes they are equally good predictors, but it is nearly redundant to use them both. Our further analysis drops weight and uses width with color and spine condition as the potential explanatory variables.

5.1.3 Stepwise Variable Selection Algorithms

Algorithms exist that select explanatory variables for a model, or delete them, in a stepwise manner. In exploratory studies, such model selection methods can be informative when used cautiously. *Backward elimination* begins with a complex model and sequentially removes

terms. In one version, at a given stage it eliminates the term in the model that has the largest P-value in the test that its parameters equal zero. The process stops when any further deletion leads to a significantly poorer fit. An alternative, *forward selection*, starts with the null model and adds terms sequentially until further additions do not improve the fit.

With either process, for categorical explanatory variables with more than two categories, the process should consider the entire variable at any stage rather than just individual indicator variables. Otherwise, the result depends on how you choose the baseline category for the indicator variables. Add or drop the entire variable rather than just one of its indicators. Also, the process should test only the highest-order terms for each variable. It is inappropriate in backward elimination, for instance, to remove a main effect term if the model contains higher-order interactions involving that term. It is not sensible to use a model with interaction but not the main effects that make up that interaction. Otherwise, the statistical significance and practical interpretation of a higher-order term depends on how the variables are coded. By including all the lower-order effects that make up an interaction, the same results occur no matter how variables are coded.

Variable selection methods need not yield a meaningful model. You should regard its results with skepticism. When you evaluate many terms, one or two that are not truly important may look impressive merely due to chance. In any case, true statistical significance is not simple to judge for effects highlighted as being the most or the least significant, and it should not be the sole criterion for whether to include a term in a model. It is sensible to include a variable that is important for the purposes of the study and report its estimated effect even if it is not statistically significant. If the variable is a potential confounder, including it in the model may help to reduce bias in estimating relevant effects of key explanatory variables and may make it possible to compare results with other studies where the effect is significant, perhaps because of a larger sample size. On the other hand, with a very large n, sometimes a term might be statistically significant but not practically significant. You might then exclude it from the model because the simpler model is easier to interpret — for example, when the term is a complex interaction.

5.1.4 Purposeful Selection of Explanatory Variables

Strategies exist for selecting explanatory variables that take into account issues such as the study goals, relative statistical significance, multicollinearity, and potential confounding. A *purposeful selection* process proposed by Hosmer et al. (2013, Chapter 4) uses several steps to build a model. In abbreviated form:

1. Construct an initial main-effects model using explanatory variables that include the known important variables and others that show *any* evidence of being relevant when used as sole predictors (e.g., having P-value < 0.2).

2. Conduct backward elimination, keeping a variable if it is either significant at a somewhat more stringent level or shows evidence of being a relevant confounder, in the sense that the estimated effect of a key variable changes substantially when it is removed.

3. Add to the model any variables that were not included in step 1 but that are significant when adjusting for the variables in the model after step 2, since a variable may not be significantly associated with y but may make an important contribution in the presence of other variables.

4. Check for plausible interactions among variables in the model after step 3, using significance tests at conventional levels such as 0.05.

5. Conduct follow-up diagnostic investigations such as those presented in Section 5.2.

5.1.5 Example: Variable Selection for Horseshoe Crabs

We illustrate a purposeful selection process using Table 5.1. It summarizes results of fitting and comparing several logistic regression models for the presence of horseshoe crab satellites, with crab width (W), color (C), and spine condition (S) as potential explanatory variables, regarding C and S as factors. For simplicity, the table symbolizes models by their terms. For instance, $(C + W)$ denotes the model with color and width main effects, whereas $(C + W + C*W)$ denotes the model with those main effects plus their interaction.

Table 5.1 Results of fitting several logistic regression models to predict horseshoe crab satellites.

Model	Explanatory Variables	Deviance	*df*	AIC	Models Compared	Deviance Difference
1	None	225.8	172	227.8		
2	C	212.1	169	220.1	(2) - (1)	13.7 ($df = 3$)
3	S	223.2	170	229.2	(3) - (1)	2.5 ($df = 2$)
4	W	194.5	171	198.5	(4) - (1)	31.3 ($df = 1$)
5	$C + W$	187.5	168	197.5	(5) - (2)	24.6 ($df = 1$)
					(5) - (4)	7.0 ($df = 3$)
6	$C + W + S$	186.6	166	200.6	(6) - (5)	0.9 ($df = 2$)
7	$C + W + C*W$	183.1	165	199.1	(7) - (5)	4.4 ($df = 3$)

Note: $C =$ color, $S =$ spine condition, $W =$ width.

At step 1, we compare the null model (model 1 in Table 5.1) to models that have color, spine condition, and width as sole predictors (models 2, 3, and 4). The likelihood-ratio statistics equal the difference in deviances between the null model and each model. These show that color and width are statistically significant. Therefore, after step 1, the purposeful selection process includes color and width as initial explanatory variables, which is model 5 in Table 5.1. At step 2, backward elimination compares model 5 to models 2 and 4 that remove weight or remove color. A large increase in deviance (24.6 on $df = 1$) results from removing width, and a moderate increase (7.0 on $df = 3$) results from removing color ($P = 0.07$), so we leave them both. At step 3, we compare model 5 to model 6 that adds spine condition, which was not one of the initially chosen variables. The decrease in deviance (0.9 on $df = 2$) is not significant, so we keep width and color as the only predictors. At step 4, model 7 adds the interaction between width and color. We implement this by adding three cross-product terms between width and the indicator variables for color. The deviance decreases by 4.4 on $df = 3$, which is not significantly better (P-value $= 0.22$). The final model developed for diagnostic investigation has solely color and width explanatory variables as main effects. That model has fit as shown in Section 4.4.1.

In this process, we treated color and spine condition as qualitative factors. If we instead treat them in a quantitative manner with equally spaced scores, as we did in Section 4.4.3, purposeful selection also yields a final model with color and width predictors. The analysis in Section 4.4.3 also indicated that the special case of model 5 that has a single indicator variable for color, equaling 1 for dark crabs and 0 otherwise, fits essentially as well.

5.1.6 AIC and the Bias/Variance Tradeoff

In selecting a model through some model-building process, you should not conclude that you have found the "correct" one. Any model is a simplification of reality. For example, you should not expect crab width to have *exactly* a linear effect on the logit, with exactly the same slope for each crab color. But, a simple model that fits adequately has the advantages of model parsimony. If a model has relatively little bias, describing reality well, it provides good estimates of outcome probabilities and of odds ratios that describe effects.

In selecting a model for any statistical analysis, a fundamental tradeoff occurs between bias and variance. Consider two models, one of which is simpler than the other. The simpler model has more potential bias, such as a true probability differing more greatly from the value corresponding to fitting the model to the population. However, the simpler model has the advantage that its smaller number of parameters results in a smaller variance for estimating characteristics of interest. Therefore, it is not necessarily better to select the more complex model, even though it has a smaller deviance and perhaps fits better in terms of statistical significance.

Because of the bias/variance tradeoff and the advantages of parsimonious models and analyses, other criteria besides significance tests can help select a good model. The best known is the *Akaike information criterion* (AIC). It judges a model by how close its fitted values tend to be to the true expected values, as summarized by a certain expected distance between the two. The optimal model, which tends to have its fitted values closest to the true outcome probabilities, is the one that minimizes

$$\text{AIC} = -2(\text{log likelihood}) + 2(\text{number of parameters in model}).$$

Because smaller AIC is better, the AIC penalizes a model for having many parameters. Even though a simpler model necessarily has smaller log likelihood (and larger deviance) than a more complex model, the simpler model may well provide better estimates of the true expected values.

For example, with a quantitative predictor x, the model $\text{logit}[\pi(x)] = \alpha + \beta_1 x + \beta_2 x^2 + \cdots + \beta_{10} x^{10}$ contains the model $\text{logit}[\pi(x)] = \alpha + \beta_1 x$ as a special case. Therefore, it fits the sample data better. Suppose, however, that the true relationship is approximately linear and $\{\beta_2, \ldots, \beta_{10}\}$ for the polynomial model that fits best to the population are all very close to 0. Because of random variability, $\{\hat{\beta}_2, \ldots, \hat{\beta}_{10}\}$ may be far from 0 and the 0 estimates of those parameters for the simpler model may be better. The fit of the more complex model with its many bends will largely reflect artifacts of the particular sample and the simpler model will yield better estimates of $\pi(x)$ at most x values.

We illustrate the AIC using the models that Table 5.1 lists. For the model $C + W$, the following R output reports a -2(log-likelihood) value of 187.46:

```
------------------------------------------------------------------------------
> fit <- glm(y ~ width + factor(color), family=binomial, data=Crabs)
> -2*logLik(fit)
'log Lik.' 187.457 (df=5)
> AIC(fit) # adds 2(number of parameters) = 2(5) = 10 to -2*logLik(fit)
[1] 197.457
------------------------------------------------------------------------------
```

The model has 5 parameters — an intercept and a width effect and three coefficients of indicator variables for color. Thus, AIC = 187.46 + 2(5) = 197.46. Of models using some or all of the three basic explanatory variables, AIC is smallest for that model.

AIC can also be the basis of stepwise model selection. For instance, the next R output uses AIC to select a model in a backward manner. We start with all four potential explanatory variables as main effects. At each step we remove the variable so that AIC decreases the most, until we get to the stage in which AIC increases if we remove any other variables:

```
------------------------------------------------------------------------------
> fit <- glm(y ~ weight + width + factor(color) + factor(spine),
+            family=binomial, data=Crabs)
> library(MASS)
> stepAIC(fit) # stepwise backward selection using AIC
Start:  AIC=201.2
y ~ weight + width + factor(color) + factor(spine)
Step:  AIC=198.21
y ~ weight + width + factor(color)
Step:  AIC=197.46
y ~ width + factor(color) # AIC now increases if width or color removed
------------------------------------------------------------------------------
```

Alternatively, we can search for the subset that has the smallest AIC when used as explanatory variables in a logistic model. This results in the same model. Here, we show this approach with color and spine condition treated as quantitative:

```
------------------------------------------------------------------------------
> attach(Crabs) # need response variable in last column of data file
> Crabs2 <- data.frame(weight, width, color, spine, y) # y in last column
> library(leaps)
> library(bestglm)
> bestglm(Crabs2, family=binomial, IC="AIC") # can also use IC="BIC"
Best Model:
              Estimate  Std. Error   z value     Pr(>|z|)
(Intercept)  -10.07084     2.80683  -3.58797  3.33261e-04
width          0.45831     0.10402   4.40606  1.05270e-05
color         -0.50905     0.22368  -2.27576  2.28602e-02
------------------------------------------------------------------------------
```

Alternatively, we can use *model averaging* to combine results from several models that seem similar in quality according to an index such as AIC. This is beyond our scope. See Burnham and Anderson (2003) for details about AIC and related indices and model averaging. They recommend using a bias-corrected version AIC_c when the number of parameters is relatively large. An alternative criterion, called *BIC*, replaces 2 by $\log(n)$ in the penalty multiple of the number of parameters. Designed to select the "correct" model as n gets very large, this index penalizes more strongly for having many parameters. Minimizing it can yield a more parsimonious model. When n is quite large, this is useful, because models can

then more easily have effects that are statistically significant but not practically significant. For the horseshoe crab data, all these criteria select the same model.

5.2 MODEL CHECKING

Any particular logistic regression model may or may not fit the data well. We next present ways to check the model fit.

5.2.1 Goodness of Fit: Model Comparison Using the Deviance

One way to detect lack of fit uses a likelihood-ratio test to compare the tentatively chosen model to more complex ones. Models with multiple predictors would consider interaction terms. If more complex models do not fit better, this provides some assurance that a chosen model is adequate. For example, the purposeful selection process summarized by Table 5.1 compared the deviances for the model having width and color main effects and the model that adds an interaction. The interaction fit is not significantly better. If a more complex model does fit better, you should evaluate whether the improvement is practically as well as statistically significant. For example, suppose that you get a small P-value by adding a set of interaction terms, but the multiple correlation measure R presented in Section 4.6.3 increases only from 0.455 to 0.463 and AIC decreases only slightly. Then, you may be better off without the extra terms, because of the simpler interpretation. Statistical significance need not imply practical significance, especially with large n.

A more general way to detect lack of fit searches for *any* way the model fails. A goodness-of-fit test compares the model fitted values to the observed responses. This approach regards the data as representing the fit of the most complex model possible — the saturated model (Section 3.4.3), which has a separate parameter for each observation. In testing fit, we test whether *all* parameters that are in the saturated model but not in the working model equal zero. The likelihood-ratio statistic for this test is the residual deviance of the model. In certain cases, this test statistic has a large-sample chi-squared null distribution.

When the explanatory variables are solely categorical, we can summarize the data by the grouped-data format of counts in a contingency table. For the n_i subjects at setting i of the explanatory variables, let $\hat{\pi}_i$ denote the estimated probability of success for the model fit. Then, the estimated binomial mean $n_i\hat{\pi}_i$ is the fitted number of successes. The fitted number of failures is $n_i(1-\hat{\pi}_i)$. The residual deviance then has the form introduced in equation (2.4), namely

$$G^2 = 2\sum \text{observed}\,[\log(\text{observed}/\text{fitted})]$$

for all the cells in that table.[2] Under the hypothesis that the model truly holds, the deviance has a large-sample chi-squared null distribution with degrees of freedom that are the residual *df* for the model. Large deviance values provide evidence of lack of fit. The P-value is the right-tail probability. The chi-squared approximation is adequate if all or nearly all of the fitted values are at least about 5.

[2] A corresponding Pearson chi-squared statistic is $X^2 = \sum(\text{observed} - \text{fitted})^2/\text{fitted}$.

5.2.2 Example: Goodness of Fit for Marijuana Use Survey

We illustrate by checking the model used in Section 4.3.2 for the data about marijuana use, gender, and race. Here is edited R output about the fit:

```
------------------------------------------------------------------------------
> Marijuana
   race gender yes  no
1 white female 420 620
2 white   male 483 579
3 other female  25  55
4 other   male  32  62
> fit <- glm(yes/(yes+no) ~ gender + race, weights=yes+no, family=binomial,
+            data=Marijuana)
> fit$deviance; fit$df.residual
[1] 0.05798  # residual deviance goodness-of-fit statistic
[1] 1        # residual df
> 1 - pchisq(fit$deviance, fit$df.residual)
[1] 0.80972 # P-value for deviance goodness-of-fit test
> fitted(fit)
      1        2        3        4
0.40453  0.45413  0.30357  0.34802 # estimated prob's of marijuana use
> fit.yes <- n*fitted(fit); fit.no <- n*(1 - fitted(fit))
> attach(Marijuana)
> data.frame(race, gender, yes, fit.yes, no, fit.no)
   race  gender  yes    fit.yes   no     fit.no
1 white  female  420  420.71429  620  619.28571
2 white    male  483  482.28571  579  579.71429
3 other  female   25   24.28571   55   55.71429
4 other    male   32   32.71429   62   61.28571
------------------------------------------------------------------------------
```

The residual deviance is 0.058. The model applies to four binomial observations, one at each of the four combinations of gender and race. The model has three parameters, so the residual $df = 4 - 3 = 1$. The small deviance suggests that the model fits decently ($P = 0.81$). However, are the fitted values large enough to trust a chi-squared approximation? For the model fit, white female students had estimated probability 0.405 of having used marijuana. Since the sample had 1040 white females, the fitted number of white female marijuana users is $1040(0.405) = 420.71$ and the fitted number of nonusers is $1040(1 - 0.405) = 619.29$. The output shows that the other six fitted values are also relatively large.

5.2.3 Goodness of Fit: Grouped versus Ungrouped Data and Continuous Predictors

Section 3.2.5 noted that with discrete explanatory variables, the data file can have the form of *ungrouped* or *grouped* data. The ungrouped data are the raw 0 and 1 observations. The grouped data are the totals of successes and failures at each combination of the predictor

values. Although the ML estimates of parameters are the same for either form of data, the deviance is not. The deviance goodness-of-fit test only makes sense for the grouped data, as the large-sample theory applies for contingency tables.

When any explanatory variables are continuous, the data file is ungrouped. Therefore, when calculated for logistic regression models having any continuous or nearly-continuous explanatory variables, the residual deviance does *not* have an approximate chi-squared distribution. How can we check the adequacy of a model for such data? Ways exist that discretize the predictors or the estimated probabilities.[3] From a scientific perspective, however, it is more informative to merely compare the working model to more complex models to judge whether additional terms provide a practically improved fit. A large deviance merely indicates *some* lack of fit, but provides no insight about its nature. Comparing a model to a more complex model, on the other hand, indicates whether lack of fit exists of a particular type. For either approach, when the fit is poor, diagnostic measures describe the influence of individual observations on the model fit and can highlight reasons for the inadequacy.

5.2.4 Residuals for Logistic Models with Categorical Predictors

With categorical explanatory variables and a grouped data file, residuals can compare observed and fitted counts. Let y_i denote the number of *successes* for the n_i trials at setting i of the explanatory variables. For a GLM with binomial random component, the *standardized residual* (Section 3.4.5) comparing y_i to its fitted value $n_i\hat{\pi}_i$ divides $(y_i - n_i\hat{\pi}_i)$ by its SE,

$$\text{Standardized residual} = \frac{y_i - n_i\hat{\pi}_i}{SE}.$$

The standardized residual has an approximate standard normal distribution when the model holds. An absolute value larger than roughly 2 or 3 provides evidence of lack of fit. Such an observation may be different from the others in some identifiable way, perhaps suggesting another variable that should be in the model. Sometimes substantive results are the same when the model is re-fitted without the observation, in which case the observation is unusual but not influential.

When fitted values are very small, residuals have limited meaning. For ungrouped binary data, each $n_i = 1$ and y_i can equal only 0 or 1, and a raw residual $(y_i - \hat{\pi}_i)$ can assume only two values. Plots of raw residuals are also then uninformative. When data can have grouped-data form, residuals are more relevant for the grouped data than for individual subjects.

5.2.5 Example: Graduate Admissions at University of Florida

Table 5.2 refers to graduate school applications to the 23 departments in the College of Liberal Arts and Sciences at the University of Florida, during a recent academic year. It cross-classifies whether the applicant was admitted (y), the applicant's gender, and the applicant's department. For the n_{ik} applications by gender i in department k, let y_{ik} denote the number admitted and let π_{ik} denote the probability of admission. We treat $\{Y_{ik}\}$ as independent binomial variates for $\{n_{ik}\}$ trials with success probabilities $\{\pi_{ik}\}$.

[3] One such approximation that some software reports is the *Hosmer–Lemeshow test*.

Table 5.2 Whether admitted (yes or no) to the graduate school at the University of Florida by gender and department, showing standardized residuals for the *yes* cell for females, for the model with no gender effect.

	Females		Males		Std. Res.		Females		Males		Std. Res.
Dept	Yes	No	Yes	No	(Fem,Yes)	Dept	Yes	No	Yes	No	(Fem,Yes)
anth	32	81	21	41	-0.76	ling	21	10	7	8	1.37
astr	6	0	3	8	2.87	math	25	18	31	37	1.29
chem	12	43	34	110	-0.27	phil	3	0	9	6	1.34
clas	3	1	4	0	-1.07	phys	10	11	25	53	1.32
comm	52	149	5	10	-0.63	poli	25	34	39	49	-0.23
comp	8	7	6	12	1.16	psyc	2	123	4	41	-2.27
engl	35	100	30	112	0.94	reli	3	3	0	2	1.26
geog	9	1	11	11	2.17	roma	29	13	6	3	0.14
geol	6	3	15	6	-0.26	soci	16	33	7	17	0.30
germ	17	0	4	1	1.89	stat	23	9	36	14	-0.01
hist	9	9	21	19	-0.18	zool	4	62	10	54	-1.76
lati	26	7	25	16	1.65						

Note: Thanks to Dr. James Booth for these data, in `Admissions` data file at text website.

Other things being equal, one would hope that the admissions decision is independent of gender. The model with no gender effect, given department, is

$$\text{logit}(\pi_{ik}) = \alpha + \beta_k.$$

However, the model may be inadequate, perhaps because a gender effect exists in some departments. The model's residual deviance is 44.7 with $df = 23$, which indicates that it fits rather poorly (P-value $= 0.004$).

Table 5.2 also reports standardized residuals for the number of females who were admitted for this model. For instance, the Astronomy department admitted 6 females, which was 2.87 standard errors higher than predicted by the model. Each department has $df = 1$ (the df for independence in a 2×2 table) and only a single nonredundant standardized residual. The standardized residuals are identical in absolute value for males and females but of different sign. Astronomy admitted 3 males, and their standardized residual was -2.87, that is, 2.87 standard errors lower than predicted. Departments with large standardized residuals are responsible for the lack of fit. Significantly more females were admitted than the model predicts in the Astronomy and Geography departments, and significantly fewer were admitted in the Psychology department. Without these three departments, the residual deviance drops to 24.4, with $df = 20$.

The model that also has a gender effect does not provide an improved fit (deviance $=$ 42.4, $df = 22$), because the departments just described have associations in different directions and of greater magnitude than other departments. Interestingly, this model has an ML estimate of 1.19 for the conditional odds ratio between Y and gender: The estimated odds of admission were *higher* for females than males, given department. By contrast, the marginal table collapsed over department has a sample odds ratio of 0.94, the overall odds of admission being *lower* for females. This illustrates Simpson's paradox (Section 2.7.3) because the estimated conditional association has a different direction than the marginal association.

5.2.6 Standardized versus Pearson and Deviance Residuals

When you request a summary of a `glm` fit in `R`, the output first shows *deviance residuals* (Section 3.4.5). For instance, for the logistic model for marijuana use, for which we showed output in Section 5.2.2:

```
------------------------------------------------------------------------------
> fit <- glm(yes/(yes+no) ~ gender + race, weights=yes+no, family=binomial,
+            data=Marijuana)
> summary(fit)
Deviance Residuals:
       1         2         3         4
-0.04513   0.04402   0.17321  -0.15493
------------------------------------------------------------------------------
```

We do not show deviance residuals in the outputs in this book. Unlike standardized residuals, deviance residuals and Pearson residuals do not appropriately recognize parameter redundancies. The residual $df = 1$ for the logistic model fitted to the four binomials on marijuana use, but the output shows 4 distinct deviance residuals rather than 1. Here is what we get with standardized (Pearson) residuals, compared to Pearson residuals, deviance residuals, and standardized deviance residuals:

```
------------------------------------------------------------------------------
> cbind(rstandard(fit,type="pearson"), residuals(fit,type="pearson"),
>       residuals(fit,type="deviance"), rstandard(fit,type="deviance"))
> # standardized, Pearson, deviance  std. dev. residuals
1    -0.24096    -0.04513  -0.04513  -0.24098 # white females
2     0.24096     0.04402   0.04402   0.24095 # white males
3     0.24096     0.17368   0.17321   0.24031 # other females
4    -0.24096    -0.15466  -0.15493  -0.24138 # other males
------------------------------------------------------------------------------
```

More sensibly, since residual $df = 1$, only one absolute value occurs (0.24096) for the standardized residuals.[4]

5.2.7 Influence Diagnostics for Logistic Regression

As in ordinary regression modeling, one observation may have much more influence than the others in determining the parameter estimates. The fit could be quite different if it were deleted. Whenever a residual indicates that a model fits an observation poorly, it can be informative to delete the observation and re-fit the model to the remaining ones. However, a single observation can have a more exorbitant influence in ordinary regression than in logistic regression, because ordinary regression has no bound on the distance of y_i from its expected value. Thus, influence measures with logistic regression are most useful when

[4] When $df = 1$, its square is the Pearson goodness-of-fit statistic X^2 (see footnote 2 in Section 5.2.1). Standardized deviance residuals are approximately, but not exactly, equal.

considered for observations in grouped-data files, for which each observation is a binomial response for a set of subjects all having the same predictor values.

Several diagnostics describe various aspects of influence. Many of them relate to the effect on certain characteristics of removing the observation from the data set. Influence diagnostics for each observation include:

1. The standardized residual and measures using it (such as *Cook's distance*) that describe how each observation contributes to lack of fit.
2. For each model parameter, the change in the parameter estimate when the observation is deleted. This change, divided by its standard error, is called *Dfbeta*.
3. The decrease in the deviance when the observation is deleted.

For each diagnostic, the larger the value, the greater the influence.

5.2.8 Example: Heart Disease and Blood Pressure

Table 5.3 comes from an analysis of data from the Framingham study, a longitudinal study of male subjects in Framingham, Massachusetts. In this analysis, men aged 40–59 were classified on $x =$ blood pressure and $y =$ whether heart disease developed during a six-year follow-up period. Let π_i be the probability of heart disease for blood pressure category i with score x_i that is the midpoint for an interval of blood pressure. The table shows the fit for the linear logistic model, which is $\text{logit}(\hat{\pi}_i) = -6.0820 + 0.0243x_i$.

Table 5.3 Diagnostic measures for logistic regression model fitted to heart disease data.

Blood Pressure	Sample Size	Observed Disease	Fitted Disease	Standardized Residual	*Dfbeta*	Deviance Decrease
111.5	156	3	5.2	−1.11	0.49	1.39
121.5	252	17	10.6	2.37	−1.14	5.04
131.5	284	12	15.1	−0.95	0.33	0.94
141.5	271	16	18.1	−0.57	0.08	0.34
151.5	139	12	11.6	0.13	0.01	0.02
161.5	85	8	8.9	−0.33	−0.07	0.11
176.5	99	16	14.2	0.65	0.40	0.42
191.5	43	8	8.4	−0.18	−0.12	0.03

Source: J. Cornfield, *Fed. Proc. 21, Suppl.* **11**: 58–61 (1962). Data are in `HeartBP` data file at text website.

Table 5.3 also reports standardized residuals, an approximation[5] for the *Dfbeta* measure for the coefficient of blood pressure, and the decrease in the deviance. All their values show that deleting the second observation has the greatest effect. One relatively large diagnostic is not surprising, however. With many observations, a small percentage may be large merely by chance.

For these data, the residual deviance of 5.91 with $df = 6$ does not indicate lack of fit. In analyzing diagnostics, we should be cautious about attributing patterns to what might

[5] Reported by SAS (PROC LOGISTIC). Influence measures are available in `R` with the `influence.measures` function in the `stats` package.

be a chance variation from a model. Also, these deletion diagnostics all relate to removing an entire binomial sample at a blood pressure level instead of removing a single subject's binary observation. Such subject-level deletions have very little effect for this model.

Another useful graphical display for showing lack of fit compares observed and fitted proportions by plotting them against each other, or by plotting both of them against explanatory variables. For the linear logistic model, Figure 5.1 plots both the observed proportions and the fitted probabilities of heart disease against blood pressure. The fit seems decent.

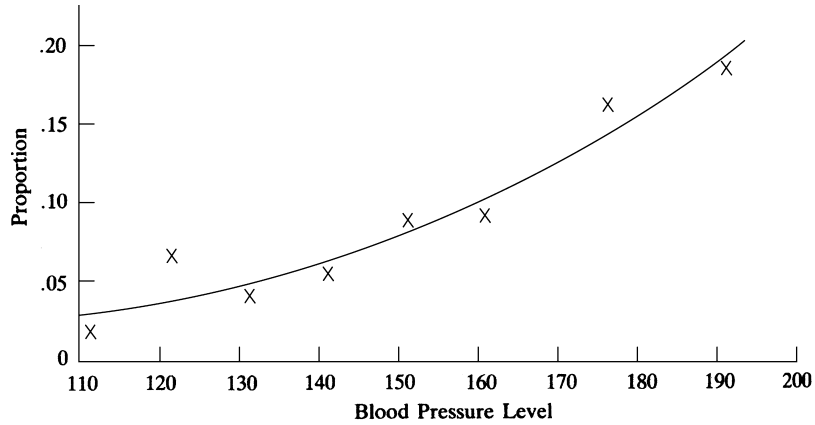


Figure 5.1 Observed proportion (x) and fitted probability of heart disease (curve) for the linear logistic model.

5.3 INFINITE ESTIMATES IN LOGISTIC REGRESSION

For logistic regression modeling, the Fisher scoring algorithm for fitting GLMs (Section 3.5.1) usually converges quickly to the correct values. However, when the successes separate from the failures in the range of values of the explanatory variables over which they each occur, the ML estimates are infinite or do not even exist.

5.3.1 Complete and Quasi-Complete Separation: Perfect Discrimination

With a single explanatory variable x, the ML estimate for its effect is infinite when the x values having $y = 0$ are completely below or completely above those having $y = 1$. Figure 5.2 illustrates, showing data with $y = 0$ at $x = 10, 20, 30, 40$ and $y = 1$ at $x = 60, 70, 80, 90$. The explanatory variable values exhibit *complete separation*, as we can split the x values into two parts, with $y = 0$ whenever x is relatively small and $y = 1$ whenever x is relatively large. There is then *perfect discrimination*: You can predict the sample y values perfectly by knowing whether x is below or above 50. A slightly weaker condition, called *quasi-complete separation*, occurs when observations of both type occur at the point of separation in the x values. This would happen if we also had two observations at $x = 50$, with $y = 1$ for one and $y = 0$ for the other.

An ideal (perfect) fit for the data in Figure 5.2 has $\hat{\pi} = 0$ for $x \leq 40$ and $\hat{\pi} = 1$ for $x \geq 60$. A sequence of logistic curves that approaches this ideal results from letting $\hat{\beta}$ increase without limit, with $\hat{\alpha} = -50\hat{\beta}$. (This $\hat{\alpha}$ value yields $\hat{\pi} = 0.50$ at $x = 50$.) In fact, the

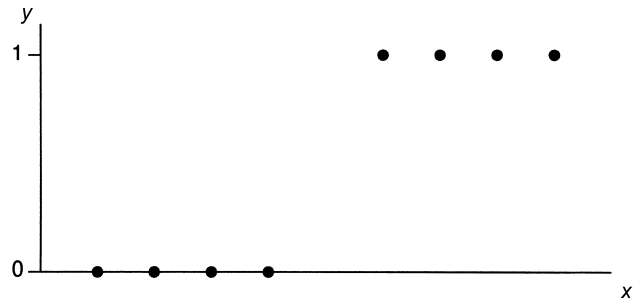


Figure 5.2 Complete separation resulting in perfect discrimination and an infinite ML logistic regression parameter estimate.

likelihood function keeps increasing as we take such a sequence, and the ML estimate of β is $\hat{\beta} = \infty$.

With several explanatory variables, suppose a plane can pass through the multidimensional space for the data such that on one side of that plane, $y = 0$ for all observations, whereas on the other side of the plane, $y = 1$ always. Then again at least one ML parameter estimate will be infinite. In practice, most software reports an inappropriate estimate when an ML estimate is truly infinite.[6] After a certain number of cycles of the iterative fitting process, the log likelihood looks flat at the working estimate, and convergence criteria are satisfied for finding the maximum. Because the log likelihood is so flat and because standard errors of ML estimates increase as the curvature decreases, software then reports huge standard errors.

5.3.2 Example: Infinite Estimate for Toy Example

For the data in Figure 5.2, here is what R yields for the logistic regression fit, after 25 iterations of the Fisher scoring fitting process:

```
------------------------------------------------------------------------------
> x <- c(10, 20, 30, 40, 60, 70, 80, 90); y <- c(0, 0, 0, 0, 1, 1, 1, 1)
> fit <- glm(y ~ x, family = binomial)
Warning message: glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(fit)
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -118.158  296046.187        0         1
x               2.363    5805.939        0         1 # P-value for Wald test
---                                                    # of H0: beta=0
    Null deviance: 1.1090e+01  on 7  degrees of freedom
Residual deviance: 2.1827e-10  on 6  degrees of freedom # res. deviance = 0
                                                         # perfect fit
Number of Fisher Scoring iterations: 25 # very slow convergence
> logLik(fit) # maximized log-likelihood = 0, so maximized likelihood = 1
'log Lik.' -1.09134e-10 (df=2)
```

[6] Some software can detect separation, such as the `brglm2` package in R for bias reduction, as shown in the output in Section 5.3.4.

```
> library(car)
> Anova(fit)
  LR Chisq  Df  Pr(>Chisq) # P-value for likelihood-ratio test of beta=0
x    11.09   1    0.000868 # more sensible than P-value of 1 for Wald test
> library(profileModel) # ordinary confint function fails for infinite est's
> confintModel(fit, objective="ordinaryDeviance", method="zoom",
+              endpoint.tolerance = 1e-08)
         Lower    Upper
x      0.05876      Inf # 95% profile likelihood CI for beta
-------------------------------------------------------------------------------
```

Although $\hat{\beta} = \infty$, R reports $\hat{\beta} = 2.36$ with $SE = 5805.94$.

As a consequence of the huge SE value, the Wald statistic $z = \hat{\beta}/SE$ is worthless. In this example, $z = 0$ and the P-value is 1. By contrast, even with a truly infinite ML estimate, the likelihood-ratio test is valid. The difference between the null deviance (i.e., for the model forcing $\beta = 0$) of 11.09 and the residual deviance of 0 is 11.09 with $df = 1$. This test has P-value = 0.0009 and yields very strong evidence of an effect. A 95% profile likelihood confidence interval for β is (0.059, ∞), corresponding to a 1-unit multiplicative effect on the odds of at least $e^{0.059} = 1.06$. The infinite upper endpoint reflects that the likelihood function keeps increasing all the way out to $\hat{\beta} = \infty$.

5.3.3 Sparse Data and Infinite Effects with Categorical Predictors

Infinite estimates also can occur with categorical explanatory variables. With a single binary explanatory variable x, such as in comparing two groups on y, we can display the data as counts in a 2×2 contingency table. The logistic regression model has an indicator variable for x, and $\hat{\beta}$ is the sample log odds ratio in the 2×2 table. When one of the four cell counts is 0, $\hat{\beta} = \pm\infty$. Such a table exhibits quasi-complete separation, with one group having outcomes of both types but the other group having only failures or only successes. Complete separation occurs when one group has only successes and the other group has only failures, in which case also $\hat{\beta} = \pm\infty$. When neither group has successes or neither group has failures, the sample odds ratio has form 0/0 and $\hat{\beta}$ does not exist.

With several categorical explanatory variables, a multiway contingency table can display the data. When the table has a large number of cells, most cell counts are usually small and many may equal 0. Such contingency tables are said to be *sparse*. A cell with a count of 0 is said to be *empty*. Although empty, in the population the cell's true probability is almost always positive. That is, it is theoretically possible to have observations in the cell, and a positive count would occur if the sample size were sufficiently large. To emphasize this, such an empty cell is often called a *sampling zero*.

Depending on the model, sampling zeroes may or may not cause ML estimates of model parameters to be infinite. When any marginal counts corresponding to terms in a model equal zero, infinite estimates occur for that term. For instance, when the model has p main effect factors, an infinite ML estimate occurs when a two-way marginal table relating Y to a factor has a zero count. If the model also has an x_1x_2 interaction term, an effect will be infinite if a cell count in the three-way table relating Y to x_1 and x_2 is zero, which is not unusual. When all cell counts are positive, all parameter estimates are necessarily finite.

5.3.4 Example: Risk Factors for Endometrial Cancer Grade

A study[7] about endometrial cancer with 79 patients analyzed how y = histology grade (0 = low, 1 = high) relates to three risk factors: x_1 = neovasculation (1 = present, 0 = absent), x_2 = pulsatility index of arteria uterina (ranging from 0 to 49), and x_3 = endometrium height (ranging from 0.27 to 3.61). Table 5.4 shows some of the data.

Table 5.4 Part of endometrial cancer data set[a].

HG	NV	PI	EH	HG	NV	PI	EH	HG	NV	PI	EH
0	0	13	1.64	0	0	16	2.26	0	0	8	3.14
...											
1	1	21	0.98	1	0	5	0.35	1	1	19	1.02

[a] HG = histology grade, NV = neovasculation, PI = pulsatility index, EH = endometrium height.
Source: Data courtesy of Ella Asseryanis, Georg Heinze, and Michael Schemper. Complete data ($n = 79$) in `Endometrial` file at text website.

For these data, we fitted the main effects model

$$\text{logit}[P(Y = 1)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

When $x_1 = 0$, both response outcomes occur, but for all 13 patients having $x_1 = 1$, the outcome is $y_i = 1$. Therefore, the data exhibit quasi-complete separation. The ML estimate $\hat{\beta}_1 = \infty$, but here is what we get in R:

```
------------------------------------------------------------------------------
> Endo <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Endometrial.dat",
+                    header=TRUE)
> Endo
   NV PI   EH HG # HG is histology grade response variable
1   0 13 1.64  0
2   0 16 2.26  0
...
79  0 33 0.85  1
> xtabs(~NV + HG, data=Endo) # quasi-complete separation:
    HG                     # when NV=1, no HG=0 cases occur
NV    0   1
  0  49  17
  1   0  13
> fit <- glm(HG ~ NV + PI + EH, family=binomial, data=Endo)
> summary(fit)
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   4.305       1.637    2.629   0.0086
NV           18.186    1715.751    0.011   0.9915 # true estimate = infinity
PI           -0.042       0.044   -0.952   0.3413
EH           -2.903       0.846   -3.433   0.0006
---
     Null deviance: 104.903  on 78  degrees of freedom
Residual deviance:  55.393  on 75  degrees of freedom
```

[7] From G. Heinze and M. Schemper, *Statist. Medic.* **21**: 2409–2419 (2002).

```
> logLik(fit) # not exactly 0 because separation is quasi, not complete
'log Lik.' -27.69663 (df=4)
> library(car)
> Anova(fit)
   LR Chisq  Df  Pr(>Chisq) # likelihood-ratio tests
NV   9.3576   1     0.00222 # compare to Wald P-value = 0.9915 for NV effect
PI   0.9851   1     0.32093
EH  19.7606   1   8.777e-06

> library(profileModel)  # ordinary confint function fails for infinite est.
> confintModel(fit, objective="ordinaryDeviance", method="zoom",
+               endpoint.tolerance = 1e-08)
         Lower     Upper
NV     1.28411       Inf # 95% profile likelihood CI for beta1
PI    -0.13708   0.03818
EH    -4.78591  -1.43639

> library(brglm2) # contains method for detecting infinite estimates
> glm(HG ~ NV + PI + EH, family=binomial,data=Endo,method="detectSeparation")
Separation: TRUE
(Intercept)    NV  PI  EH
          0   Inf   0   0 # 0 denotes finite est., Inf denotes infinite est.
-------------------------------------------------------------------------------
```

Despite $\hat{\beta}_1 = \infty$, inference is possible about β_1. The likelihood-ratio statistic for H_0: $\beta_1 = 0$ equals 9.36 with $df = 1$ and has P-value $= 0.002$. The 95% profile likelihood confidence interval for β_1 is (1.28, ∞). We can conclude that $\beta_1 > 0$ (despite what the Wald P-value shows on the R output!) and that the effect is substantial. Figure 5.3 portrays how this works, showing a normalized profile log-likelihood function for β_1 that keeps increasing toward 0 as β_1 increases indefinitely toward ∞. The values of $\beta_1 > 1.28$ have $L(\beta_1)$ close enough to its maximum to be considered plausible values for that parameter.

The other ML estimates are not affected by the quasi-complete separation. In fact, most of the predictive power is provided by EH. The correlation between the observed y values and the fitted values (Section 4.6.3) is 0.745 for the full model and 0.692 for the model with EH as the sole predictor; the areas under the ROC curves (Section 4.6.2) are 0.907 and 0.895. More complex models, not presented here, do not provide an improved fit.

When infinite ML estimates exist, we normally expect the true effects in the population to be finite. Two methods presented in the next section produce finite estimates when ML estimates are infinite.

5.4 BAYESIAN INFERENCE, PENALIZED LIKELIHOOD, AND CONDITIONAL LIKELIHOOD FOR LOGISTIC REGRESSION *

For fitting logistic regression models and making inferences, alternatives exist to the ML frequentist approach presented so far. The alternatives include Bayesian modeling and modified types of likelihood-based analyses that yield estimators that can have less bias and take finite value when ML estimates are infinite.

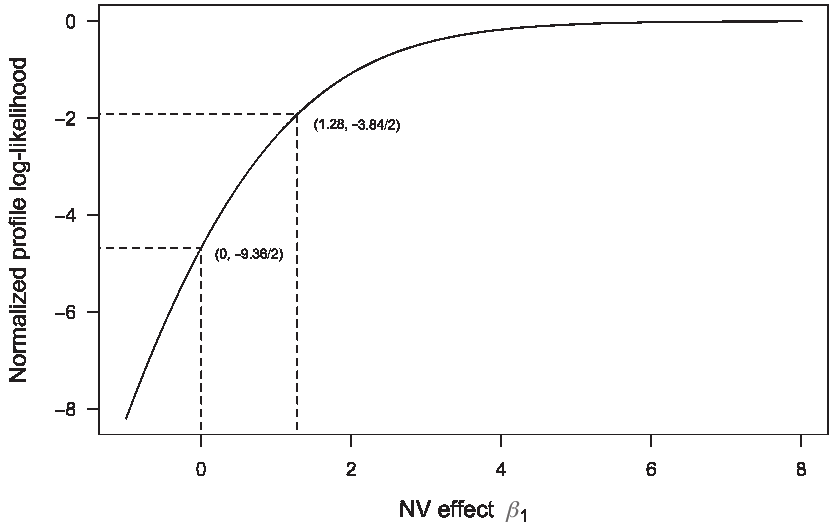


Figure 5.3 Normalized profile log-likelihood function $L(\beta_1) - L(\hat{\beta}_1)$ for NV effect in the main-effects logistic model. Double the log-likelihood increases by 9.36 (the likelihood-ratio test statistic) between $\beta_1 = 0$ and $\hat{\beta}_1 = \infty$ and by 3.84 (the test statistic value that yields chi-squared P-value $=$ 0.05) between $\beta_1 = 1.28$ and $\hat{\beta}_1 = \infty$. *Source*: Figure constructed by Alessandra Brazzale with `cond` R package for higher-order likelihood-based conditional inference for logistic models.

5.4.1 Bayesian Modeling: Specification of Prior Distributions

For Bayesian inference with logistic regression modeling, it is common to treat the $\{\beta_j\}$ as independent normal random variables, with means of 0. Most data analysts take the standard deviation for the normal distributions to be large, so that the prior distribution has relatively little effect on the results, which are then similar substantively to those with the frequentist approach.

The next example illustrates the potentially large impact on the results of the choice of variability for the prior distribution.

5.4.2 Example: Risk Factors for Endometrial Cancer Revisited

Section 5.3.4 described a study about endometrial cancer that analyzed $y =$ histology grade of 79 cases, with the explanatory variables neovasculation, pulsatility index of arteria uterina, and endometrium height. For the main-effects model, all 13 patients having neovasculation present had $y = 1$. Therefore, quasi-complete separation occurs, and the ML estimate $\hat{\beta}_1 = \infty$. To make the effect magnitudes comparable for the two quantitative explanatory variables, as explained in Section 4.5.3, we can use standardized variables. Here are `R` results for ML, continuing the analysis from Section 5.3.4.

```
------------------------------------------------------------------------------
> Endo$PI2 <- scale(Endo$PI); Endo$EH2 <- scale(Endo$EH) # standardizes
> Endo$NV2 <- Endo$NV - 0.5 # rescale useful for later Bayesian analysis
> fit.ML <- glm(HG ~ NV2 + PI2 + EH2, family=binomial, data=Endo)
```

```
> summary(fit.ML) # ML estimate of NV2 effect is actually infinite
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)    7.8411    857.8755    0.009   0.9927
NV2           18.1856   1715.7509    0.011   0.9915
PI2           -0.4217      0.4432   -0.952   0.3413
EH2           -1.9219      0.5599   -3.433   0.0006
-------------------------------------------------------------------------
```

In our Bayesian analyses, we use $N(\mu, \sigma^2)$ prior distributions for the $\{\beta_j\}$. To reflect a lack of prior belief about the direction of the effects, we took each $\mu = 0.0$. Instead of the usual (0, 1) coding for the indicator variable x_1, we let it take values -0.5 and 0.5. The prior distribution is then symmetric in the sense that the logits for each neovasculation group have the same prior variability as well as the same prior means, yet β_1 still has the usual interpretation of a conditional log odds ratio.

For these data, because the log likelihood is so flat in the β_1 dimension, posterior means for β_1 can be highly dependent on the prior σ value. To reflect a lack of information about the effect sizes, we first took the prior distributions to be quite diffuse, with $\sigma = 10$. Here are R results:

```
-----------------------------------------------------------------------------
> library(MCMCpack) # b0 = prior mean, B0 = prior precision = 1/variance
> fitBayes <- MCMClogit(HG ~ NV2 + PI2 + EH2, mcmc=10000000, b0=0, B0=0.01,
+                       data=Endo) # prior var. = 1/0.01 = 100, std dev = 10
> summary(fitBayes)
1. Empirical mean and standard deviation: # posterior distribution
               Mean     SD
(Intercept)   3.215  2.560
NV2           9.120  5.097
PI2          -0.473  0.454
EH2          -2.138  0.593
2. Quantiles for each variable:
               2.5%     25%     50%     75%   97.5%
(Intercept)  -0.342   1.271   2.722   4.687   9.346
NV2           2.109   5.234   8.128  12.048  21.343
PI2          -1.414  -0.767  -0.455  -0.159   0.366
EH2          -3.403  -2.515  -2.101  -1.722  -1.082
> mean(fitBayes[,2] < 0) # probability below 0 for 2nd model parameter (NV2)
[1] 0.000223
-----------------------------------------------------------------------------
```

Table 5.5 shows posterior means and standard deviations and the 95% equal-tail posterior interval for β_1, based on an MCMC process with 10,000,000 iterations. With such a long process, the Monte Carlo standard errors for the approximations to the Bayes estimates were negligible — about 0.005 for the neovasculation effect and much less for the others. Based on the posterior mean, the Bayesian estimated odds of the higher-grade histology when neovasculation is present are $\exp(9.12) = 9136$ times the estimated odds when neovasculation is absent. The 95% equal-tail posterior interval for β_1 of $(2.1, 21.3)$ provides the inference that $\beta_1 > 0$ and that the effect is strong. The estimated effect size is imprecise

Table 5.5 Results of Bayesian and frequentist fitting of models to the endometrial cancer data-set of Table 5.4 .

Analysis	$\hat{\beta}_1$ (SD)	Interval[a]	$\hat{\beta}_2$ (SD)	$\hat{\beta}_3$ (SD)
ML	∞ (—)	(1.3, ∞)	-0.42 (0.44)	-1.92 (0.56)
Bayes, $\sigma = 10$	9.12 (5.10)	(2.1, 21.3)	-0.47 (0.45)	-2.14 (0.59)
Bayes, $\sigma = 1$	1.65 (0.69)	(0.3, 3.0)	-0.22 (0.33)	-1.77 (0.43)

[a] Profile-likelihood interval for ML and equal-tail posterior interval for Bayes.

because of the flat log likelihood and the relatively flat prior distribution, but the interval does not go all the way out to ∞ as it does with a profile likelihood interval in the frequentist approach. Inferences about β_2 and β_3 were substantively the same as with the ML frequentist analysis, also shown in the table for comparison.

Corresponding to the frequentist P-value for testing H_0: $\beta_1 = 0$ against H_a: $\beta_1 > 0$, the Bayesian approach provides the posterior probability that $\beta_1 \leq 0$. This is approximated as 0.0002; that is, 0.0 is the 0.0002 quantile of the posterior distribution. For this relatively flat prior distribution, this posterior tail probability is similar to the P-value of 0.001 for the one-sided frequentist likelihood-ratio test. Each provides very strong evidence that $\beta_1 > 0$.

For comparison, we performed a Bayesian analysis with a highly informative prior distribution. To reflect a prior belief that the effects are not very strong, we took $\sigma = 1.0$ with $\mu = 0$. Then nearly all the prior probability for the conditional odds ratio $\exp(\beta_1)$ falls between $\exp(-3.0) = 0.05$ and $\exp(3.0) = 20$. As Table 5.5 shows, results were quite different from the ML frequentist analysis or the Bayesian analysis with $\sigma = 10$. Because $y_i = 1$ for all 13 patients having $x_{i1} = 1$, the frequentist approach tells us we cannot rule out any extremely large positive value for β_1. By contrast, if we had strong prior beliefs that $|\beta_1| < 3$, then even with these sample results the Bayesian posterior interval infers an upper bound of about 3 for β_1.

5.4.3 Penalized Likelihood Reduces Bias in Logistic Regression

In fitting GLMs, *regularization methods* modify ML to give sensible estimates in potentially unstable situations, such as for highly sparse contingency tables. One way to do this adds a term to the log-likelihood function such that maximizing it smooths the ordinary ML estimate. For a model with log-likelihood function $L(\boldsymbol{\beta})$ for the model parameters $\boldsymbol{\beta}$, the method maximizes

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) - s(\boldsymbol{\beta}),$$

where $s(\cdot)$ is a function such that $s(\boldsymbol{\beta})$ decreases as elements of $\boldsymbol{\beta}$ are smoother in some sense, such as uniformly closer to 0. This smoothing method, referred to as *penalized likelihood*, shrinks the ML estimates toward 0.

The statistician David Firth showed[8] that the ML estimator in logistic regression is biased away from 0 and proposed a penalized-likelihood correction that reduces the bias. Maximizing the Firth penalized log-likelihood function yields an estimate that always exists and

[8] D. Firth, *Biometrika* **80**: 27–38 (1993).

is unique. The penalized likelihood estimates are often more believeable than ML estimates, such as when ML estimates are infinite or badly affected by multicollinearity.

5.4.4 Example: Risk Factors for Endometrial Cancer Revisited

Sections 5.3.4 and 5.4.2 described a study about endometrial cancer that modeled $y =$ histology grade with the explanatory variables neovasculation, pulsatility index, and endometrium height. Quasi-complete separation occurs in terms of neovasculation, for which the ML estimate is infinite. Table 5.5 showed Bayes estimates, which shrink $\hat{\beta}_1$ from ∞ to 9.12 for quite diffuse normal priors ($\sigma = 10$) and to 1.65 for very informative priors ($\sigma = 1$). Here is what we get with Firth's penalized likelihood approach, which you can also compare with ordinary ML estimates in Table 5.5:

```
------------------------------------------------------------------------------
> library(logistf) # can implement Firth's penalized likelihood method
> fit.penalized <- logistf(HG ~ NV2 + PI2 + EH2, family=binomial, data=Endo)
> summary(fit.penalized)
Confidence intervals and p-values by Profile Likelihood
               coef  se(coef)  lower 0.95 upper 0.95   Chisq          p
(Intercept)  0.3080    0.8006     -0.9755     2.7888   0.169  6.810e-01
NV2          2.9293    1.5508      0.6097     7.8546   6.798  9.124e-03
PI2         -0.3474    0.3957     -1.2443     0.4045   0.747  3.875e-01
EH2         -1.7243    0.5138     -2.8903    -0.8162  17.759  2.507e-05
------------------------------------------------------------------------------
```

The maximum penalized-likelihood estimate for β_1 of 2.93 and the 95% profile penalized-likelihood confidence interval of (0.61, 7.85) shrink the ML estimate $\hat{\beta}_1 = \infty$ and the ordinary profile likelihood interval of (1.28, ∞) considerably toward 0. Results for the other estimates do not change as much.

5.4.5 Conditional Likelihood and Conditional Logistic Regression

ML estimators of logistic regression model parameters usually perform well when the sample size n is large compared with the number of parameters. When n is small or when the number of parameters is very large, improved inference results when using *conditional maximum likelihood*.

This inferential approach deals with the primary parameters of interest using a *conditional likelihood* function that eliminates the other parameters — so-called *nuisance parameters*. This likelihood function is called "conditional" because the way to eliminate the nuisance parameters is to find a probability distribution of the data that conditions on the potential samples that provide the same information about the nuisance parameters that occurs in the observed sample. This conditional distribution and its conditional likelihood function then depend only on the parameters of interest. Using that conditional likelihood function, the usual types of inference methods apply, such as likelihood-ratio tests and confidence intervals.

One situation in which conditional likelihood methods are especially useful is when the model has a huge number of parameters, perhaps a similar number of parameters as observations, but the main focus is on a single effect. For example, some models that handle

repeated observations on subjects have a term in the model for each subject. In a longitudinal study, suppose that y_{it} and x_{it} are the values of the response variable and an explanatory variable for person i at time t. In the model

$$\text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta x_{it},$$

each person has their own intercept (α_i). This permits variability in the success probability among people at a particular value of the explanatory variable, perhaps due to variability in other explanatory variables not in the model. For n people, the model has $n + 1$ parameters. Conditional likelihood methods can eliminate the $\{\alpha_i\}$ parameters before we conduct inference about β. Section 8.2 uses such a model and a conditional analysis in the contexts of matched-pairs and case-control studies.

5.4.6 Conditional Logistic Regression and Exact Tests for Contingency Tables

With categorical explanatory variables, the conditional likelihood approach can generate *exact* sampling distributions. Then, probabilities such as P-values can use the exact distributions rather than normal or chi-squared approximations. Such exact distributions are possible because the conditioning eliminates the nuisance parameters from the likelihood function. Methods that find and use the exact distribution are computationally intensive, but some software is available.[9]

To illustrate, when x is a binary indicator, the logistic regression model $\text{logit}[P(Y = 1)] = \alpha + \beta x$ applies to 2×2 contingency tables of counts $\{n_{ij}\}$ for which the two columns are the levels of Y. The usual sampling model treats the responses on Y in the two rows as independent binomial samples. The row totals, which are the numbers of trials for those binomial variates, are naturally fixed. When we test H_0: $\beta = 0$, α is a nuisance parameter. It refers to the relative number of outcomes of $y = 1$ and $y = 0$, which are the column totals. We can eliminate α from the likelihood function by conditioning also on the column totals, which are the information in the data about α. Fixing both sets of marginal totals yields a hypergeometric distribution for n_{11}, for which the probabilities do not depend on α. The resulting exact test of H_0: $\beta = 0$ is Fisher's exact test (Section 2.6.1).

For multiway contingency tables, we can also use conditional logistic regression to conduct exact inference for a parameter of interest, by eliminating the other model parameters from the likelihood function. This is beyond our scope. For details, see Agresti (2013, Section 7.3).

5.5 ALTERNATIVE LINK FUNCTIONS: LINEAR PROBABILITY AND PROBIT MODELS *

Although logistic regression is the most popular model for binary response variables, models with other link functions are sometimes more appropriate and can be simpler to interpret. In this section we present two alternatives — the *linear probability model* and the *probit model*.

[9] For example, SAS (PROC LOGISTIC), `StatXact` and `LogXact` (Cytel Software), and the `clogit` and `LogistiX` packages in `R`.

5.5.1 Linear Probability Model

Section 3.2.1 introduced the *linear probability model*. With multiple explanatory variables, the model has form

$$P(Y = 1) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p,$$

for which the probability of success changes linearly in each x_j. This model is a GLM with a binomial random component and identity link function.

As we have noted, this model structure generates predicted values over the entire real line, but probabilities fall between 0 and 1. This limits the scope of its applicability. Iterative algorithms for finding the ML estimates under the binomial sampling assumption fail when, at some stage, an estimated probability $\hat{P}(Y = 1) = \hat{\alpha} + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$ falls outside the 0 to 1 range for some observation. Software then usually displays an error message such as "lack of convergence." If we ignore the binary nature of Y and use ordinary regression modeling, the $\{\hat{\beta}_j\}$ are the *least squares* estimates. They are the ML estimates under the assumption of a normal distribution for Y with constant variance. These estimates exist because for a normal random variable an estimated mean of Y can be any real number and is not restricted to the 0 to 1 range. When ML fitting with the binomial assumption fails, the least squares method applied to the ungrouped data file succeeds but usually also yields some estimated probabilities outside the 0 to 1 range.

The linear probability model has the advantage that an estimated effect is simple to interpret, as a slope on the probability scale. It takes a similar value as the *average marginal effect* (Section 4.5.2) for the fit of the corresponding logistic model.

5.5.2 Example: Political Ideology and Belief in Evolution

Section 3.4.2 presented General Social Survey results on $y =$ opinion about evolution (1 = true, 0 = false) and $x =$ political ideology (1 = extremely conservative to 7 = extremely liberal). For the linear probability model, treating x as quantitative, ML fitting yields $\hat{P}(Y = 1) = 0.108 + 0.110x$. We obtain a similar fit with an ordinary linear model. Here is edited R output, for the ungrouped data file `Evolution2` at the text website:

```
------------------------------------------------------------------------------
> Evo <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Evolution2.dat",
+                   header=TRUE)
> Evo
     ideology evolved # ungrouped data file, n=1064
1           1       1
2           1       1
....
1064        7       0
> fit <- glm(evolved ~ ideology, family=quasi(link=identity,
+            variance="mu(1-mu)"), data=Evo)
> summary(fit, dispersion=1)
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   0.10844     0.03893    2.785   0.00535
ideology      0.11010     0.00897   12.273   < 2e-16
---
```

```
    Null deviance: 1469.3  on 1063  degrees of freedom
Residual deviance: 1359.5  on 1062  degrees of freedom

> fit2 <- glm(evolved ~ ideology, family=gaussian(link=identity), data=Evo)
> summary(fit2) # ordinary linear model, used if no convergence with binom.
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   0.10554     0.04272     2.47    0.0137
ideology      0.11091     0.01033    10.73    <2e-16
--------------------------------------------------------------------------------
```

The estimated probability of believing in evolution increases by 0.11 for each 1-unit change in the liberal ideology direction. This estimated effect is a simple and useful summary. The corresponding logistic model has fit logit $[\hat{P}(Y = 1)] = -1.757 + 0.494x$ and has an average marginal effect equal to 0.111. The linear probability model provides fitted probabilities of belief in evolution ranging from 0.218 for extremely conservative to 0.879 for extremely liberal. These describe well the corresponding sample proportions, as the deviance for the grouped version of the data file is 3.45 ($df = 5$).

5.5.3 Probit Model and Normal Latent Variable Model

Another model that has S-shaped curves like those of logistic regression is called the *probit model*. It has expression

$$\text{probit}[P(Y = 1)] = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{5.1}$$

The link function for the model, called the *probit link*, transforms $P(Y = 1)$ to the standard normal z-score at which the left-tail probability equals $P(Y = 1)$. For instance, probit(0.05) $= -1.645$, because 5% of the standard normal distribution falls below -1.645. Likewise, probit(0.50) $= 0$, probit(0.95) $= 1.645$, and probit(0.975) $= 1.96$.

Interpretation of parameters in probit models is simplest when we can relate the model to a corresponding normal linear model. Many binary variables can be regarded as a crude measurement of an unobservable continuous variable. For example, suppose we use a binary regression model for $y =$ political ideology, where each subject chooses *liberal* or *conservative* for the response. In practice, differences in political ideology exist among people who classify themselves in the same category. With a precise enough way to measure political ideology, it is possible to imagine an essentially continuous measurement, reflecting how people can range from extremely liberal to extremely conservative with a great number of possibilities between the two. In statistics, an unobserved variable assumed to underlie what we actually observe is called a *latent variable*.

Suppose a latent variable y^* exists such that the observed binary response $y = 0$ if $y^* \leq \tau$ and $y = 1$ if $y^* > \tau$, for some unknown threshold τ. Suppose also that the latent variable satisfies an ordinary linear model

$$y^* = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon.$$

When ϵ has a normal distribution with the same variance at all values for the explanatory variables, necessarily the probit model holds for the observed binary response. Moreover, if we scale the latent response so that var(ϵ) $= 1$, the effects in the probit model are the

same as in the latent variable model. Then we can interpret $\hat{\beta}_j$ from the probit model fit as representing the estimated change in $E(y^*)$ for a 1-unit increase in x_j, adjusting for the other explanatory variables. With arbitrary value for var(ϵ), $\hat{\beta}_j$ is the estimated number of standard deviations that the distribution of y^* shifts.

5.5.4 Example: Snoring and Heart Disease Revisited

Table 3.1 (Section 3.2.3) showed data from a study of the potential impact of snoring on the presence of heart disease. The response was binary, but one can conceive of a potential continuous measurement y^* of the degree of heart disease. With scores (0, 2, 4, 5) for snoring level, the ML fit of the probit model is

$$\text{probit}[\hat{P}(Y = 1)] = -2.061 + 0.188x.$$

For the corresponding latent variable model for y^*, a 1-unit increase in $x =$ snoring level corresponds to a 0.188 standard deviation increase in $E(y^*)$. As x increases from 0 to 5, the underlying latent distribution of heart disease shifts up by nearly a standard deviation (i.e., $5(0.188) = 0.94$).

At snoring level $x = 0$, the probit equals $-2.061 + 0.188(0) = -2.06$. The fitted probability of heart disease is the left-tail probability for the standard normal distribution at $z = -2.06$, which equals 0.020. At snoring level $x = 5$, the probit equals $-2.061 + 0.188(5) = -1.12$, which corresponds to a fitted probability of 0.131. Here is edited R output:

```
------------------------------------------------------------------------------
> Heart <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Heart.dat",
+                     header=TRUE)
> Heart
             snoring yes   no
1              never  24 1355
2         occasional  35  603
3 nearly_every_night  21  192
4        every_night  30  224
> library(dplyr)
> Heart$x <- recode(Heart$snoring, never = 0, occasional = 2,
+                   nearly_every_night = 4, every_night = 5)
> fit <- glm(yes/(yes+no) ~ x, family=binomial(link=probit),
+            weights=yes+no, data=Heart)
> summary(fit)
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -2.06055     0.07017  -29.367   < 2e-16
x             0.18777     0.02348    7.997  1.28e-15
---
    Null deviance: 65.9045  on 3  degrees of freedom
Residual deviance:  1.8716  on 2  degrees of freedom
> fitted(fit) # estimated P(Y=1) at 4 snoring levels for x
      1        2        3        4
0.01967  0.04599  0.09519  0.13100
------------------------------------------------------------------------------
```

5.5.5 Latent Variable Models Imply Binary Regression Models

Other binary regression models result from the latent variable structure when we assume other distributions for ϵ. For example, the *logistic distribution* looks similar to a normal distribution but with slightly thicker tails. The latent variable model in which ϵ has a logistic distribution implies the logit link and a logistic regression model for the observed binary response.

In practice, probit and logistic regression models provide similar fits, because normal and logistic distributions are so similar. If a logistic regression model fits well, then so does the probit model, and conversely. For the snoring and heart disease data, for example, the residual deviance is 2.81 with the logit link and 1.87 with the probit link, each with $df = 2$. Parameter estimates in probit models have smaller magnitude than those in logistic regression models. This is because their link functions transform probabilities to scores from standard versions of the normal and logistic distribution, but those two distributions have a different spread. The standard normal distribution has $\mu = 0$ and $\sigma = 1$. The standard logistic distribution has $\mu = 0$ and $\sigma = 1.8$. When both models fit well, parameter estimates in logistic regression models are approximately 1.8 times those in probit models.

5.5.6 CDFs and Shapes of Curves for Binary Regression Models

The assumed distribution for an underlying latent variable also determines the shape of the curve for $P(Y = 1)$, as explained now. For a random variable Z, the *cumulative distribution function* (*cdf*) F for Z is the function that specifies all the *cumulative probabilities*,

$$F(z) = P(Z \leq z), \ -\infty < z < \infty.$$

As z increases over its range of values, $F(z)$ increases from 0 to 1. When Z is a continuous random variable, the *cdf*, plotted as a function of z, has S-shaped appearance like those obtained with logistic regression curves. This suggests a class of models for binary responses whereby the dependence of $P(Y = 1)$ on the explanatory variables has the form

$$P(Y = 1) = F(\alpha + \beta_1 x_1 + \cdots + \beta_p x_p), \tag{5.2}$$

where F is a *cdf* for some continuous probability distribution. Then F^{-1} is the link function applied to $P(Y = 1)$ to yield the linear predictor.

When F is the *cdf* of a standard normal distribution, F^{-1} is the probit link function. That link function transforms $P(Y = 1)$ so that the curve for $P(Y = 1)$ has the appearance of the normal *cdf*, as a function of the linear predictor. With a single explanatory variable, the parameters (μ, σ) of the normal *cdf* for $P(Y = 1)$ relate to the parameters in the probit model by $\mu = -\alpha/\beta$ and $\sigma = 1/|\beta|$. Each choice of α and of $\beta > 0$ corresponds to a different normal distribution.

For the snoring and heart disease data, $\text{probit}[\hat{P}(Y = 1)] = -2.061 + 0.188x$. This probit fit corresponds to a normal *cdf* for $P(Y = 1)$ having $\hat{\mu} = -\hat{\alpha}/\hat{\beta} = 2.061/0.188 = 11.0$ and $\hat{\sigma} = 1/|\hat{\beta}| = 1/0.188 = 5.3$. The estimated probability of heart disease equals 1/2 at snoring level $x = 11.0$. Since the snoring level is restricted to the range 0 to 5 for these data, well below 11, the fitted probabilities over this range are quite small.

The logistic regression curve has form (5.2) with F as the *cdf* of a standard logistic distribution. The curve for $P(Y = 1)$ then has the shape of the *cdf* of a logistic distribution.

5.6 SAMPLE SIZE AND POWER FOR LOGISTIC REGRESSION *

The major aim of many studies is to determine whether a particular variable has an effect on a binary response variable. The study design should determine the sample size needed to provide a high probability of detecting an effect of a practically significant size.

5.6.1 Sample Size for Comparing Two Proportions

For a study designed to compare two groups, consider the hypothesis that the group *success* probabilities π_1 and π_2 are identical. We could conduct a test for the 2×2 table that cross-classifies group by response, rejecting H_0: $\pi_1 = \pi_2$ if the P-value $\leq \alpha$ for some fixed α. To determine sample size, we must specify the probability β of failing to detect a difference between π_1 and π_2 of some fixed size considered to be practically important. For this size of effect, β is the probability of failing to reject H_0 at the α level. Then, $\alpha = P$(Type I error) and $\beta = P$(Type II error). The *power* of the test equals $1 - \beta$.

A study using equal group sample sizes requires approximately

$$n_1 = n_2 = (z_{\alpha/2} + z_\beta)^2[\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)]/(\pi_1 - \pi_2)^2.$$

This formula requires values for π_1, π_2, α, and β. It is approximate but adequate for most practical work, because conjectures for values of π_1 and π_2 are usually very approximate.

For testing H_0: $\pi_1 = \pi_2$ at the 0.05 level with P(Type II error) $= 0.10$, suppose that $\pi_1 = 0.20$ and $\pi_2 = 0.30$. Then $\alpha = 0.05$, $\beta = 0.10$, $z_{\alpha/2} = z_{.025} = 1.96$, $z_\beta = z_{.10} = 1.28$, and we require

$$n_1 = n_2 = (1.96 + 1.28)^2[(0.2)(0.8) + (0.3)(0.7)]/(0.2 - 0.3)^2 = 389.$$

This formula also provides the sample sizes needed for a comparable confidence interval for $\pi_1 - \pi_2$. Then, $1 - \alpha$ is the confidence level for the interval and β equals the probability that the interval indicates a plausible lack of effect, in the sense that it contains 0. Therefore, we need about 400 subjects in each group for a 95% confidence interval for $\pi_1 - \pi_2$ to have only a 0.10 chance of containing 0 when actually $\pi_1 = 0.20$ and $\pi_2 = 0.30$.

5.6.2 Sample Size in Logistic Regression Modeling

Next, consider testing H_0: $\beta_1 = 0$ for the logistic regression model[10]

$$\text{logit}[P(Y = 1)] = \beta_0 + \beta_1 x.$$

When x is binary, the model compares success probabilities for two groups, for which we just discussed sample size determination. When x is continuous, the sample size needed to

[10] We use β_0 for the intercept so as not to confuse it with $\alpha = P$(Type I error).

obtain a certain power is difficult to determine, because it depends on the distribution of the x values.

Under the assumption that X is random and has a normal distribution, the required sample size depends on the probability of success $\bar{\pi}$ at the mean of X and the odds ratio θ comparing $\bar{\pi}$ to the probability of success one standard deviation above the mean. Let $\lambda = \log(\theta)$. For a one-sided test,[11]

$$n = [z_\alpha + z_\beta \exp(-\lambda^2/4)]^2(1 + 2\bar{\pi}\delta)/(\bar{\pi}\lambda^2),$$

where

$$\delta = [1 + (1 + \lambda^2)\exp(5\lambda^2/4)]/[1 + \exp(-\lambda^2/4)].$$

The value n decreases as $\bar{\pi}$ gets closer to 0.50 and as $|\lambda|$ gets farther from the null value of 0.

A multiple logistic regression model requires larger n to detect a partial effect of the same size. Let R denote the multiple correlation between the explanatory variable x_j of primary interest and the others in the model. In this formula for n, we divide by $(1 - R^2)$, with $\bar{\pi}$ being the probability at the mean value of all the explanatory variables and θ being the effect of x_j at the mean of the others. However, this result is of limited use, because even if $R = 0$, an effect in a multiple logistic regression model changes in magnitude when variables are added to a model.[12]

These formulas provide, at best, rough ballpark indications of sample size. In most applications, we have only a crude guess for $\bar{\pi}$, θ, and R, and the explanatory variable of main interest may be far from normally distributed.

5.6.3 Example: Modeling the Probability of Heart Disease

A research study plans to model how the probability of severe heart disease depends on $x =$ cholesterol level for a middle-aged population. Previous studies have suggested that $\bar{\pi}$ is about 0.08. Suppose the investigators want the test of H_0: $\beta_1 = 0$ against H_a: $\beta_1 > 0$ to be sensitive to a 50% increase, for a standard deviation increase in cholesterol. The odds of severe heart disease at the mean cholesterol level equal $0.08/0.92 = 0.087$, and the odds one standard deviation above the mean equal $0.12/0.88 = 0.136$. The odds ratio equals $\theta = 0.136/0.087 = 1.57$, from which $\lambda = \log(1.57) = 0.450$ and $\delta = 1.306$. For $\beta =$ P(Type II error) $= 0.10$ in an $\alpha = 0.05$-level test, $z_{.05} = 1.645$, $z_{.10} = 1.28$, and the study needs $n = 612$.

EXERCISES

5.1 For the horseshoe crabs data file, fit a model using weight and width as explanatory variables for the probability of a satellite.

a. Conduct a likelihood-ratio test of H_0: $\beta_1 = \beta_2 = 0$. Interpret.

[11] Due to F.Y. Hsieh, *Statist. Medic.*, **8**: 795–802 (1989).

[12] For example, see the article by L. Robinson and N. Jewell in *Intern. Statist. Rev.* **58**: 227–240 (1991).

b. Conduct separate likelihood-ratio tests for the partial effects of each variable. Why does neither test show evidence of an effect when the test in (**a**) shows very strong evidence?

c. Use purposeful selection or AIC to build a model when weight and the spine condition and color factors are the potential explanatory variables.

5.2 Table 7.8 in Chapter 7 shows data from the `Substance2` data file at the text website. Create a new data file from which you can build a logistic regression model for these data, treating marijuana use as the response variable and alcohol use, cigarette use, gender, and race as explanatory variables. Prepare a short report summarizing a model selection process, with edited software output as an appendix.

5.3 The `Crabs2` data file at the text website shows several variables that may be associated with $y =$ whether a female horseshoe crab is monandrous (eggs fertilized by a single male crab) or polyandrous (eggs fertilized by multiple males). Using *year* of observation, *Fcolor* = the female crab's color (1 = dark, 3 = medium, 5 = light), *Fsurf* = her surface condition (values 1, 2, 3, 4, 5 with lower values representing worse), *FCW* = female's carapace width, *AMCW* = attached male's carapace width, *AMcolor* = attached male's color, and *AMsurf* = attached male's surface condition, conduct a logistic model-building process. Prepare a report summarizing this process, with edited software output as an appendix. Interpret results for your chosen model.

5.4 The `Students` data file at the text website shows responses of a class of social science graduate students at the University of Florida to a questionnaire that asked about *gender* (1 = female, 0 = male), *age*, *hsgpa* = high school GPA (on a four-point scale), *cogpa* = college GPA, *dhome* = distance (in miles) of the campus from your home town, *dres* = distance (in miles) of the classroom from your current residence, *tv* = average number of hours per week that you watch TV, *sport* = average number of hours per week that you participate in sports or have other physical exercise, *news* = number of times a week you read a newspaper, *aids* = number of people you know who have died from AIDS or who are HIV+, *veg* = whether you are a vegetarian (1 = yes, 0 = no), *affil* = political affiliation (1 = Democrat, 2 = Republican, 3 = Independent), *ideol* = political ideology (1 = very liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conservative), *relig* = how often you attend religious services (0 = never, 1 = occasionally, 2 = most weeks, 3 = every week), *abor* = opinion about whether abortion should be legal in the first three months of pregnancy (1 = yes, 0 = no), *affirm* = support affirmative action (1 = yes, 0 = no), and *life* = belief in life after death (1 = yes, 2 = no, 3 = undecided).

a. Show all steps of a model-selection method such as purposeful selection for choosing a model for predicting *abor*, when the potential explanatory variables are *ideol*, *relig*, *news*, *hsgpa*, and *gender*.

b. Using an automated tool such as the `stepAIC` or `bestglm` function in `R`, construct a model to predict *abor*, selecting from the 14 binary and quantitative variables in the data file as explanatory variables.

c. With $y =$ *veg* and the 14 binary and quantitative variables in the data file as explanatory variables, show that the likelihood-ratio test of H_0: $\beta_1 = \cdots =$

$\beta_{14} = 0$ has P-value < 0.001, yet forward selection using Wald tests with 0.05 criterion selects the null model. Explain how this could happen.

5.5 Exercise 4.12 introduced four scales of the Myers–Briggs personality test. Table 5.6 shows SAS output for fitting a model using the four scales as predictors of whether a subject drinks alcohol frequently.

a. Conduct a model goodness-of-fit test, and interpret. If you were to simplify the model by removing a predictor, which would you remove? Why?

b. Software reports AIC values of 642.1 for the model with the four main effects and the six interaction terms, 637.5 for the model with only the four binary main effect terms, and 648.8 for the model with no predictors. According to this criterion, which model is preferred? Explain the rational for using AIC.

c. Using the `MBTI` data file at the website `www.stat.ufl.edu/~aa/intro-cda/data`, use model-building methods to select a model for this alcohol response variable.

Table 5.6 SAS output for fitting model to Myers–Briggs personality scales data of Exercise 4.12.

Criterion		DF	Value		
Deviance		11	11.1491		
Parameter	Estimate	Standard Error	Like-ratio 95% Conf Limits		Chi-Square
Intercept	-2.4668	0.2429	-2.9617	-2.0078	103.10
EI e	0.5550	0.2170	0.1314	0.9843	6.54
SN s	-0.4292	0.2340	-0.8843	0.0353	3.36
TF t	0.6873	0.2206	0.2549	1.1219	9.71
JP j	-0.2022	0.2266	-0.6477	0.2426	0.80

5.6 Refer to the previous exercise. The data file also shows responses on whether a person smokes frequently. Software reports model -2 log-likelihood values of 1130.23 with only an intercept term, 1124.86 with also the main effect predictors, and 1119.87 with also all the two-factor interactions.

a. Write the model for each case and show that the numbers of parameters are 1, 5, and 11.

b. Find AIC values. Which of the three models is preferable?

5.7 For data introduced in Exercise 4.10 about AIDS symptoms, AZT use, and race, here is some `R` output:

```
> fit <- glm(yes/(yes+no) ~ azt + race, weights=yes+no, family=binomial,
+            data=AIDS)
> summary(fit)
Deviance Residuals:
       1        2        3        4
-0.5547   0.4253   0.7035  -0.6326
```

```
              Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -1.07357     0.26294   -4.083  4.45e-05
aztyes        -0.71946     0.27898   -2.579   0.00991
racewhite      0.05548     0.28861    0.192   0.84755
---
    Null deviance: 8.3499  on 3  degrees of freedom
Residual deviance: 1.3835  on 1  degrees of freedom
> 1 - pchisq(1.3835, 1)
[1] 0.23950
> cbind(AIDS$azt, AIDS$race, fitted(fit), rstandard(fit,type="pearson"),
        residuals(fit,type="pearson"), residuals(fit,type="deviance"))
  [,1] [,2]    [,3]     [,4]     [,5]     [,6]
1    2    2  0.1496  -1.1794  -0.5447  -0.5547 # azt=yes, race=white
2    1    2  0.2654   1.1794   0.4282   0.4253 # azt=no,  race=white
3    2    1  0.1427   1.1794   0.7239   0.7035 # azt=yes, race=black
4    1    1  0.2547  -1.1794  -0.6220  -0.6326 # azt=no,  race=black
------------------------------------------------------------------------------
```

a. Test the model goodness of fit and interpret the result.

b. Explain how the relative sizes of the fitted values reflect the results of the individual tests for the AZT effect and the race effect.

c. The display shows the standardized residuals, Pearson residuals, and deviance residuals. Explain advantages of using standardized residuals rather than the others.

5.8 Refer to Table 2.9 on death penalty decisions. Fit a logistic model with the two race predictors. Conduct a residual analysis and interpret.

5.9 Table 5.7 shows a $2 \times 2 \times 6$ contingency table for $y =$ whether admitted to graduate school at the University of California, Berkeley, for fall 1973, by gender of applicant for the six largest graduate departments.

a. Fit the logistic model that has department as the sole explanatory variable for y. Use the standardized residuals to describe the lack of fit.

Table 5.7 Data for Exercise 5.9 on admissions to Berkeley.

	Admitted, Male		Admitted, Female	
Department	Yes	No	Yes	No
1	512	313	89	19
2	353	207	17	8
3	120	205	202	391
4	138	279	131	244
5	53	138	94	299
6	22	351	24	317
Total	1198	1493	557	1278

Note: Based on data in P. Bickel *et al.*, *Science* **187**: 398–403 (1975).

b. When we add a gender effect, the estimated conditional odds ratio between admissions and gender (1 = male, 0 = female) is 0.90. The marginal table, collapsed over department, has odds ratio 1.84. Explain what causes these associations to differ so much.

5.10 The `Lungs` data file at the text website[13] summarizes eight studies in China about smoking and lung cancer. Analyze these data and prepare a short report that summarizes your analyses and interpretations.

5.11 Refer to the model you selected in part (**a**) of Exercise 5.4. Check goodness of fit. Can you conduct a residual analysis with this data file? Explain.

5.12 Suppose $y = 0$ at $x = 0, 10, 20, 30$ and $y = 1$ at $x = 70, 80, 90, 100$.

a. Explain intuitively why $\hat{\beta} = \infty$ for the model, $\text{logit}[P(Y = 1)] = \alpha + \beta x$. Report $\hat{\beta}$ and its SE for the software you use.

b. Add two observations at $x = 50$, $y = 1$ for one and $y = 0$ for the other. Report $\hat{\beta}$ and its SE. Do you think these are correct? Why? What happens if you replace the two observations by $y = 1$ at $x = 49.9$ and $y = 0$ at $x = 50.1$?

5.13 Refer to Exercise 5.4. With *veg* as the response variable, find a logistic model for which at least one ML effect estimate is infinite. Explain the aspect of the data file that causes this. Report and interpret results from fitting the model using either Firth's penalized logistic regression or Bayesian inference.

5.14 Table 5.8 is from a study of nonmetastatic osteosarcoma described in the *LogXact 7* manual (Cytel Software, 2005, p. 171). The response is whether the subject achieved a three-year disease-free interval.

a. Show that each explanatory variable has a significant effect when it is used as the sole predictor in logistic regression. Try to fit a main-effects model containing all three predictors. Explain why the ML estimate for the effect of lymphocytic infiltration is actually infinite.

b. Report and interpret results from fitting the main-effects model using either Firth's penalized logistic regression or Bayesian inference.

Table 5.8 Data for Exercise 5.14.

Lymphocytic Infiltration	Sex	Osteoblastic Pathology	Disease-Free	
			Yes	No
High	Female	No	3	0
High	Female	Yes	2	0
High	Male	No	4	0
High	Male	Yes	1	0
Low	Female	No	5	0
Low	Female	Yes	3	2
Low	Male	No	5	4
Low	Male	Yes	6	11

[13] Based on data in *Intern. J. Epidemiol.*, **21**: 197–201 (1992) by Z. Liu.

Table 5.9 Clinical trial relating treatment to response for five centers.

		Response	
Center	Treatment	Success	Failure
1	Active drug	0	5
	Placebo	0	9
2	Active drug	1	12
	Placebo	0	10
3	Active drug	0	7
	Placebo	0	5
4	Active drug	6	3
	Placebo	2	6
5	Active drug	5	9
	Placebo	2	12

Source: Diane Connell, Sandoz Pharmaceuticals Corp.

5.15 Table 5.9 shows results of a randomized clinical trial conducted at five centers. The purpose was to compare an active drug to placebo for treating fungal infections (1 = success, 0 = failure). For these data, let y = response, x = treatment (1 = active, 0 = placebo), and z = center.

a. For the model $\text{logit}[P(Y = 1) = \alpha + \beta x + \beta_k^z$, explain why quasi-complete separation occurs in terms of the effects of center.

b. Using a "no intercept" option so that $\{\beta_k^z\}$ refer to the individual centers rather than contrasts with a baseline center, fit the model and report $\hat{\beta}_1^z$ and $\hat{\beta}_3^z$ and their standard errors. What are the actual ML estimates?

c. The counts in the 2×2 marginal table relating treatment to response are all positive, so the empty cells do not affect the treatment estimate. Report the estimated treatment log odds ratio and show that it does not change when you delete Centers 1 and 3 from the analysis. (When a center has outcomes of only one type, it provides no information about the treatment effect.)

5.16 Refer to Exercise 4.1 on cancer remission. Table 5.10 shows output for fitting a probit model. Interpret the parameter estimates (**a**) finding the remission value at which the estimated probability of remission equals 0.50, (**b**) finding the difference between the estimated probabilities of remission at the upper and lower quartiles of the labeling index, 14 and 28, (**c**) using a corresponding normal latent variable model, (**d**) using characteristics of the normal *cdf* response curve.

Table 5.10 Table for Exercise 5.16 on probit model for cancer remission.

```
------------------------------------------------------------------------------
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -2.31777     0.76060   -3.047   0.00231
LI            0.08785     0.03293    2.668   0.00763
------------------------------------------------------------------------------
```

5.17 Refer to the `SoreThroat` data file introduced in Exercise 4.16. Fit and interpret the main effects (**a**) linear probability model, (**b**) probit model.

5.18 For the `Crabs` data file, fit the linear probability model to the probability that a female horseshoe crab with shell width x has a satellite. Is the fit adequate for large x values?

5.19 We expect success probabilities for two groups to be about 0.20 and 0.30, and we want an 80% chance of detecting a difference using a 90% confidence interval.

a. Assuming equal sample sizes, how large should they be?

b. Compare results to the sample sizes required for (i) a 90% interval with power 90%, (ii) a 95% interval with power 80%.

5.20 The width values in the `Crabs` data file have a mean of 26.3 and standard deviation of 2.1. If the true relationship is the fitted equation reported in Section 4.1.3, $\text{logit}[\pi(x)] = -12.351 + 0.497x$, about how large a sample yields P(Type II error) $= 0.10$ in an $\alpha = 0.05$-level test of H_0: $\beta = 0$ against H_a: $\beta > 0$? What assumption does this result require?

5.21 The following are true–false questions.

a. A model for a binary y has a continuous explanatory variable. If the model truly holds, the residual deviance has a distribution approaching chi-squared as n increases. It can be used to test model goodness of fit.

b. When x_1 or x_2 is the sole predictor for binary y, the likelihood-ratio test of the effect has P-value < 0.0001. When both x_1 and x_2 are in the model, it is possible that the likelihood-ratio tests for H_0: $\beta_1 = 0$ and for H_0: $\beta_2 = 0$ could both have P-values larger than 0.05.

CHAPTER 6

MULTICATEGORY LOGIT MODELS

Ordinary logistic regression applies to binary response variables. Generalizations of logistic regression apply to categorical responses that have more than two categories. Models for *nominal*-scale response variables treat the categories as unordered while models for *ordinal*-scale response variables utilize the category ordering. Explanatory variables can again be quantitative, categorical (using indicator variables), or both.

We let c denote the number of categories of the response variable Y. The response probabilities $(\pi_1, ..., \pi_c)$ at any setting for the explanatory variables satisfy $\sum_j \pi_j = 1$. As in the previous chapters, the analyses of this chapter apply when the sample consists of independent observations. When all explanatory variables are discrete, the data file can be ungrouped or can have the grouped-data form of counts in the c categories of Y at each setting of the explanatory variables. The models assume that those counts have a *multinomial* distribution — the multicategory generalization of the binomial distribution (Section 1.2.2).

6.1 BASELINE-CATEGORY LOGIT MODELS FOR NOMINAL RESPONSES

The multicategory logit model for nominal response variables simultaneously uses all pairs of categories by specifying the odds of outcome in one category instead of another. The order of listing the categories is irrelevant.

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

6.1.1 Baseline-Category Logits

The basic model formula pairs each category with a baseline category. Software usually sets the last category (c) as the baseline, in which case the *baseline-category logits* are

$$\log\left(\frac{\pi_j}{\pi_c}\right),\ j = 1, \ldots, c-1.$$

For $c = 3$, for instance, the model uses $\log(\pi_1/\pi_3)$ and $\log(\pi_2/\pi_3)$. Conditional on the response falling in category j or in category c, $\log(\pi_j/\pi_c)$ is the log odds that the response is j.

The baseline-category logit model with an explanatory variable x is

$$\log\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_j x,\ j = 1, \ldots, c-1. \tag{6.1}$$

The model has $c-1$ equations, with *separate parameters for each*. The effects vary according to the category paired with the baseline. These equations determine equations for *all* pairs of categories. When $c = 3$, for example,

$$\begin{aligned}\log\left(\frac{\pi_1}{\pi_2}\right) &= \log\left(\frac{\pi_1/\pi_3}{\pi_2/\pi_3}\right) = \log\left(\frac{\pi_1}{\pi_3}\right) - \log\left(\frac{\pi_2}{\pi_3}\right)\\ &= (\alpha_1 + \beta_1 x) - (\alpha_2 + \beta_2 x)\\ &= (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)x. \end{aligned}\tag{6.2}$$

This equation has the form $\alpha + \beta x$ with intercept parameter $\alpha = (\alpha_1 - \alpha_2)$ and with slope parameter $\beta = (\beta_1 - \beta_2)$.

With p explanatory variables, this model extends to

$$\log\left(\frac{\pi_j}{\pi_c}\right) = \alpha_j + \beta_{j1}x_1 + \beta_{j2}x_2 + \cdots + \beta_{jp}x_p,\ \ j = 1, \ldots, c-1. \tag{6.3}$$

For each explanatory variable, different logits have different effects. Unless c and p are both small, the model has a large number of parameters. Software for multicategory logit models fits all $(c-1)$ equations (6.3) *simultaneously*, using the Fisher scoring iterative algorithm. The baseline category is arbitrary and the same ML parameter estimates occur for a pair of categories no matter which baseline category you use.

6.1.2 Example: What Do Alligators Eat?

Table 6.1 comes from a study by the Florida Game and Fresh Water Fish Commission of the foods that alligators eat. For 59 alligators captured in Lake George, Florida, Table 6.1 shows the primary food type (in volume) found in the alligator's stomach. The primary food type has $c = 3$ categories: Fish, Invertebrate, and Other. The invertebrates were primarily apple snails, aquatic insects, and crayfish. The *other* category includes amphibian, mammal, plant material, stones or other debris, and reptiles (primarily turtles, although one stomach

Table 6.1 Alligator length (in meters) and primary food choice,[a] for 59 Florida alligators.

1.24 I	1.30 I	1.30 I	1.32 F	1.32 F	1.40 F	1.42 I	1.42 F
1.45 I	1.45 O	1.47 I	1.47 F	1.50 I	1.52 I	1.55 I	1.60 I
1.63 I	1.65 O	1.65 I	1.65 F	1.65 F	1.68 F	1.70 I	1.73 O
1.78 I	1.78 I	1.78 O	1.80 I	1.80 F	1.85 F	1.88 I	1.93 I
1.98 I	2.03 F	2.03 F	2.16 F	2.26 F	2.31 F	2.31 F	2.36 F
2.36 F	2.39 F	2.41 F	2.44 F	2.46 F	2.56 O	2.67 F	2.72 I
2.79 F	2.84 F	3.25 O	3.28 O	3.33 F	3.56 F	3.58 F	3.66 F
3.68 O	3.71 F	3.89 F					

[a] F = Fish, I = Invertebrates, O = Other.
Source: Thanks to M. F. Delany and Clint T. Moore for these data, in the `Alligators` data file at the text website.

contained 23 baby alligators!). The table also shows the alligator length, which varies between 1.24 and 3.89 meters.

Let Y = primary food choice and x = alligator length. Here is `R` output for fitting baseline-category logit model (6.1), with *other* as the baseline category:

```
------------------------------------------------------------------------------
> Gators <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Alligators.dat",
+                      header=TRUE)
> Gators
     x y
1  1.24 I
2  1.30 I
...
59 3.89 F
> library(VGAM) # package for multivariate GLMs, such as multinomial models
> fit <- vglm(y ~ x, family=multinomial, data=Gators) # vglm = vector GLM
> coef(fit, matrix = TRUE)
            log(mu[,1]/mu[,3])  log(mu[,2]/mu[,3])
(Intercept)             1.6177              5.6974
x                      -0.1101             -2.4654
> summary(fit)
              Estimate Std. Error  z value  Pr(>|z|)
(Intercept):1   1.6177     1.3073    1.237   0.21591
(Intercept):2   5.6974     1.7937    3.176   0.00149
x:1            -0.1101     0.5171   -0.213   0.83137 # for log[P(Y=1)/P(Y=3)]
x:2            -2.4654     0.8996   -2.741   0.00613 # for log[P(Y=2)/P(Y=3)]
---
Residual deviance: 98.3412 on 114 degrees of freedom
Reference group is level 3 of the response # reference = baseline category
------------------------------------------------------------------------------
```

The ML prediction equations for the two baseline-category logits are

$$\log(\hat{\pi}_1/\hat{\pi}_3) = 1.618 - 0.110x,$$
$$\log(\hat{\pi}_2/\hat{\pi}_3) = 5.697 - 2.465x.$$

By (6.2), the estimated log odds that the response is *fish* (category 1) rather than *invertebrate* (category 2) equals

$$\log(\hat{\pi}_1/\hat{\pi}_2) = (1.618 - 5.697) + [-0.110 - (-2.465)]x = -4.080 + 2.355x.$$

You could also find this equation by having software use invertebrates as the baseline category:

```
------------------------------------------------------------------------------
> fit2 <- vglm(y ~ x, family=multinomial(refLevel="I"), data=Gators)
> summary(fit2) # now using y=2 (I = Invertebrates) as baseline category
              Estimate Std. Error  z value  Pr(>|z|)
(Intercept):1  -4.0797     1.4686   -2.778   0.00547
(Intercept):2  -5.6974     1.7937   -3.176   0.00149
x:1             2.3553     0.8032    2.932   0.00336 # for log[P(Y=1)/P(Y=2)]
x:2             2.4654     0.8996    2.741   0.00613 # for log[P(Y=3)/P(Y=2)]
---
Residual deviance: 98.3412 on 114 degrees of freedom # same for any baseline
Reference group is level 2 of the response
> confint(fit2, method="profile") # profile likelihood confidence intervals
                 2.5 %     97.5 %
x:1            1.01118    4.19907  # beta for log[P(Y=1)/P(Y=2)]
x:2            0.87752    4.46361  # beta for log[P(Y=3)/P(Y=2)]
------------------------------------------------------------------------------
```

The estimates for a particular equation are interpreted as in binary logistic regression, conditional on the event that the outcome falls in one of those two categories. For instance, given that the primary food type is fish or invertebrate, the prediction equation for $\log(\hat{\pi}_1/\hat{\pi}_2)$ with coefficient 2.355 for size indicates that larger alligators are relatively more likely to eat fish rather than invertebrates. The estimated conditional probability that the primary food choice is fish increases in length x according to an S-shaped curve. For alligators of length $x + 1$ meters, the estimated odds that primary food type is *fish* rather than *invertebrate* equal $\exp(2.355) = 10.5$ times the estimated odds at length x meters.

The hypothesis that primary food choice is independent of alligator length is H_0: $\beta_1 = \beta_2 = 0$. The likelihood-ratio test statistic compares the null model to the working model by the difference in deviances. The next R output shows that this is $115.14 - 98.34 = 16.80$, and equivalently double the difference in log-likelihood values, $2[-49.17 - (-57.57)] = 16.80$, in each case with $df = 2$:

```
------------------------------------------------------------------------------
> fit0 <- vglm(y ~ 1, family=multinomial, data=Gators) # null model
> deviance(fit0) # deviance for working model is 98.3412
[1] 115.1419
> lrtest(fit, fit0) # lrtest function available in VGAM package for LR tests
Likelihood ratio test # test that beta_1 = beta_2 = 0
Model 1: y ~ x
Model 2: y ~ 1
   #Df   LogLik  Df   Chisq  Pr(>Chisq)
1  114  -49.171
2  116  -57.571   2  16.801     0.00022 # deviance diff. = 2(log-like. diff.)
------------------------------------------------------------------------------
```

The P-value of 0.0002 provides very strong evidence of a length effect for at least one of the logits. From the output with invertebrates (category 2) as the baseline category, the profile-likelihood confidence intervals for β_1 and β_2 reveal that both parameters are positive, with effects that are potentially quite strong. For example, given that primary food choice is fish or invertebrates, we can be 95% confident that the multiplicative effect a 1-meter increase in length on the odds of fish falls between $e^{1.011} = 2.7$ and $e^{4.199} = 66.6$. With such a relatively small n, estimated effects are very imprecise.

6.1.3 Estimating Response Probabilities

For the baseline-category logit model (6.3), the response probabilities relate to the model parameters by

$$\pi_j = \frac{e^{\alpha_j+\beta_{j1}x_1+\beta_{j2}x_2+\cdots+\beta_{jp}x_p}}{\sum_{h=1}^{c} e^{\alpha_h+\beta_{h1}x_1+\beta_{h2}x_2+\cdots+\beta_{hp}x_p}}, \; j = 1, \ldots, c.$$

The denominator is the same for each π_j, and the numerators for various j sum to the denominator, so $\sum_j \pi_j = 1$. The parameters equal zero for whichever category j is the baseline in the logit expressions. For the estimates that contrast *fish* and *invertebrate* to *other* as the baseline category (which are in the first R output panel in Section 6.1.2), the estimated probabilities of the outcomes (Fish, Invertebrate, Other) equal

$$\hat{\pi}_1 = \frac{e^{1.618-0.110x}}{1 + e^{1.618-0.110x} + e^{5.697-2.465x}},$$

$$\hat{\pi}_2 = \frac{e^{5.697-2.465x}}{1 + e^{1.618-0.110x} + e^{5.697-2.465x}},$$

$$\hat{\pi}_3 = \frac{1}{1 + e^{1.618-0.110x} + e^{5.697-2.465x}}.$$

The 1 term in each denominator and in the numerator of $\hat{\pi}_3$ represents $e^{\hat{\alpha}_3+\hat{\beta}_3 x}$ for $\hat{\alpha}_3 = \hat{\beta}_3 = 0$ with category 3 as the baseline category.

For example, for an alligator of the maximum observed length of $x = 3.89$ meters, the estimated probabilities are $(\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3) = (0.763, 0.005, 0.232)$. Very large alligators apparently prefer to eat fish. Software can display the estimated probabilities for each observation:

```
------------------------------------------------------------------------------
> fitted(fit) # estimated response probabilities for outcome categories
         F       I       O # fish, invertebrate, other
1   0.2265  0.7220  0.0515
2   0.2503  0.6925  0.0573
...
59  0.7630  0.0047  0.2323  # obs. 59 is alligator of length 3.89 meters
------------------------------------------------------------------------------
```

Figure 6.1 shows the three estimated response probabilities as a function of alligator length.

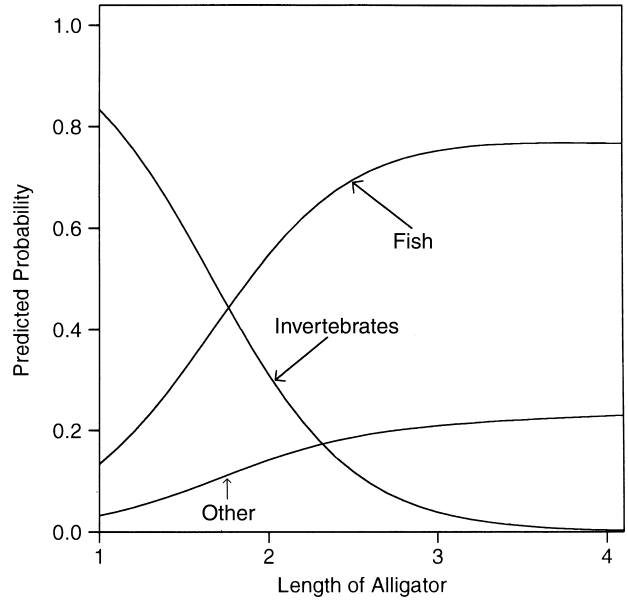


Figure 6.1 Estimated probabilities for primary food choice as a function of alligator length.

6.1.4 Checking Multinomial Model Goodness of Fit

When explanatory variables are entirely discrete, the data file can exhibit data either in *ungrouped* form or in the *grouped* contingency-table form of multinomial counts at the various combinations of categories for the explanatory variables. With the grouped-data format, when the contingency table is not too sparse, we can use the residual deviance to test the model goodness of fit.

As in the case of binary data (Section 5.2.1), the deviance has the form

$$G^2 = 2 \sum \text{observed} \, [\log(\text{observed}/\text{fitted})]$$

and is a test statistic for a global goodness-of-fit test. When most cell counts are at least about 5, then under the null hypothesis that the model holds, the deviance has an approximate chi-squared distribution. As in the binary case, this goodness-of-fit test is not valid when the model has at least one continuous explanatory variable or when the data are sparse In any case, a more informative check compares the working model to models that contain additional effects, such as additional explanatory variables or interaction terms, by comparing the deviances.

6.1.5 Example: Belief in Afterlife

Table 6.2, from a General Social Survey, has $Y =$ belief in life after death, with categories (yes, undecided, no), and explanatory variables $x_1 =$ gender and $x_2 =$ race, for which we use indicator variables. With *no* as the baseline category for Y, the model is

$$\log\left(\frac{\pi_j}{\pi_3}\right) = \alpha_j + \beta_j^G x_1 + \beta_j^R x_2, \quad j = 1, 2,$$

Table 6.2 Belief in afterlife by gender and race.

			Belief in Afterlife	
Race	Gender	Yes	Undecided	No
White	Female	371	49	74
	Male	250	45	71
Black	Female	64	9	15
	Male	25	5	13

Source: General Social Survey, `Afterlife` data file at text website.

where G and R superscripts identify the gender and race parameters. The model assumes a lack of interaction between gender and race in their effects on belief in an afterlife.

The effect parameters represent log odds ratios with the baseline category. For instance, β_1^G is the conditional log odds ratio between gender and belief in afterlife categories 1 and 3 (*yes* and *no*), given race. In the `R` output shown next, the indicator variables are (1 = male, 0 = female) for gender and (1 = white, 0 = black) for race. Since $\hat{\beta}_1^G = -0.419$, for males the estimated odds of response *yes* rather than *no* on life after death are $\exp(-0.419) = 0.66$ times those for females, adjusting for race. For whites, the estimated odds of response *yes* rather than *no* are $\exp(0.342) = 1.41$ times those for blacks, adjusting for gender.

```
------------------------------------------------------------------------------
> Afterlife <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                          Afterlife.dat", header=TRUE)
> Afterlife
   race gender yes undecided no
1 white female 371        49 74
2 white   male 250        45 71
3 black female  64         9 15
4 black   male  25         5 13
> library(VGAM)
> fit <- vglm(cbind(yes,undecided,no) ~ gender + race, family=multinomial,
+             data=Afterlife)
> summary(fit)
                 Estimate  Std. Error  z value  Pr(>|z|)
(Intercept):1      1.3016      0.2265    5.747   9.1e-09
(Intercept):2     -0.6529      0.3405   -1.918    0.0551 .
gendermale:1      -0.4186      0.1713   -2.444    0.0145
gendermale:2      -0.1051      0.2465   -0.426    0.6700
racewhite:1        0.3418      0.2370    1.442    0.1493
racewhite:2        0.2710      0.3541    0.765    0.4442
---
Residual deviance: 0.8539 on 2 degrees of freedom
------------------------------------------------------------------------------
```

For these grouped data, we can test goodness of fit using the deviance. The number of multinomial probabilities at each of the four (gender, race) combinations is $3 - 1 = 2$ (since

the probabilities must sum to 1), for a total of eight parameters. The model, considered for $j = 1$ and 2, contains six parameters. Thus, the residual deviance of 0.85 has $df = 8 - 6 = 2$. The model fits well.

The test of the gender effect has H_0: $\beta_1^G = \beta_2^G = 0$. The likelihood-ratio test shown in the next R output compares the model's deviance of 0.85 ($df = 2$) to the deviance of 8.05 ($df = 4$) that results from dropping gender from the model. The difference of deviances of $8.05 - 0.85 = 7.19$ has $df = 4 - 2 = 2$. The P-value of 0.027 shows evidence of a gender effect, with males being less likely than females to believe in an afterlife.

```
------------------------------------------------------------------------------
> fit.race <- vglm(cbind(yes,undecided,no) ~ race, family=multinomial,
+                  data=Afterlife) # removing gender from model
> deviance(fit.race)
[1] 8.04650
> lrtest(fit, fit.race) # lrtest function available in VGAM package
Likelihood ratio test
Model 1: cbind(yes, undecided, no) ~ gender + race
Model 2: cbind(yes, undecided, no) ~ race
  #Df   LogLik  Df  Chisq  Pr(>Chisq)
1   2  -19.732
2   4  -23.329   2 7.1926     0.02742 # deviance diff. = 2(log-lik. diff.)
------------------------------------------------------------------------------
```

By contrast, the effect of race is not significant: The model deleting race has deviance 2.85 ($df = 4$), which is an increase of 1.99 on $df = 2$. This lack of significance partly reflects the larger standard errors that the estimated effects of race have, due to a much greater imbalance between sample sizes in the race categories than in the gender categories.

We can also estimate probabilities for the three response categories, at each setting of gender and race:

```
------------------------------------------------------------------------------
> data.frame(Afterlife$race, Afterlife$gender, fitted(fit))
  Afterlife.race  Afterlife.gender        yes  undecided         no
1          white            female     0.7546     0.0996     0.1459
2          white              male     0.6783     0.1224     0.1993
3          black            female     0.7074     0.1002     0.1925
4          black              male     0.6222     0.1206     0.2573
------------------------------------------------------------------------------
```

According to the model fit, belief in an afterlife is most likely for white females and least likely for black males.

6.1.6 Discrete Choice Models*

The multicategory logit model is an important tool in marketing research for analyzing how people choose among a discrete set of options. For example, for individuals who recently

bought an automobile, we could model how their choice of brand depends on their annual income, size of family, attained level of education, and residence (rural or urban).

A generalization of model (6.1) allows explanatory variables to take different values for different Y categories. Such explanatory variables are *characteristics of the choices*. For example, the choice of brand of auto would likely depend on price, which varies among the brand options. The generalized model is called a *discrete-choice model*.

Discrete-choice models can also incorporate explanatory variables that are *characteristics of the chooser*, as in the examples considered so far. Thus, this model type is very general. The ordinary baseline-category logit model is a special case.

6.1.7 Example: Shopping Destination Choice*

An early use of the discrete choice model[1] analyzed how residents of Pittsburgh, Pennsylvania, chose a shopping destination. The five possible destinations were different city zones. One explanatory variable measured $S =$ shopping opportunities, defined to be the retail employment in the zone as a percentage of total retail employment in the region. The other explanatory variable was $P =$ price of the trip, defined from a separate analysis using auto in-vehicle time and auto operating cost.

Both explanatory variables are characteristics of the choices. The ML fit of the discrete-choice model yields the prediction equation for pairs of city zones a and b,

$$\log(\hat{\pi}_a/\hat{\pi}_b) = -1.06(P_a - P_b) + 0.84(S_a - S_b)$$

with standard errors 0.28 for the price of trip coefficient and 0.23 for the shopping opportunity coefficient. Not surprisingly, a destination is relatively more attractive as the trip price decreases and as the shopping opportunity increases.

For this model structure, the odds of choosing a over b do not depend on the other alternatives in the choice set or on their values for the explanatory variables. This property of the model is called *independence from irrelevant alternatives*. This may be unrealistic in applications with alternatives that are not clearly distinct. For further details about discrete choice models, see Train (2009).

6.2 CUMULATIVE LOGIT MODELS FOR ORDINAL RESPONSES

When response categories are ordered, logits can utilize the ordering. This results in models that have fewer parameters and potentially greater power and simpler interpretation than baseline-category logit models.

For the response variable Y, the cumulative probability for outcome category j is

$$P(Y \le j) = \pi_1 + \cdots + \pi_j, \;\; j = 1, \ldots, c.$$

The cumulative probabilities reflect the ordering, with

$$P(Y \le 1) \le P(Y \le 2) \le \cdots \le P(Y \le c) = 1.$$

[1] By D. McFadden, in *Frontiers in Econometrics*, ed. P. Zarembka, Academic Press (1974).

The logits of the cumulative probabilities, called *cumulative logits*, are

$$\text{logit}[P(Y \le j)] = \log\left[\frac{P(Y \le j)}{1 - P(Y \le j)}\right] = \log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_c}\right),$$

for $j = 1, \ldots, c-1$. For $c = 3$, for example, the two cumulative logits are

$$\text{logit}[P(Y \le 1)] = \log\left(\frac{\pi_1}{\pi_2 + \pi_3}\right) \quad \text{and} \quad \text{logit}[P(Y \le 2)] = \log\left(\frac{\pi_1 + \pi_2}{\pi_3}\right).$$

Cumulative logits and models for them do not use $P(Y \le c)$, because it necessarily equals 1.

6.2.1 Cumulative Logit Models with Proportional Odds

A model for cumulative logit j looks like a binary logistic regression model in which categories 1 to j combine to form one category and categories $j+1$ to c form the other. For an explanatory variable x, the model

$$\text{logit}[P(Y \le j)] = \alpha_j + \beta x, \quad j = 1, \ldots, c-1, \tag{6.4}$$

has parameter β describing the effect of x on the log odds of response in category j or below. In this formula, β does not have a j subscript. The model assumes that the effect of x is identical for all $c-1$ cumulative logits. When this model fits well, it requires only a single parameter to describe the effect of x. By contrast, the baseline-category logit model requires $c-1$ parameters, one for each logit.

Figure 6.2 depicts this model for a four-category response and quantitative x. Each cumulative probability has it own curve, describing its change as a function of x. The curve for $P(Y \le j)$ looks like a logistic regression curve for a binary response with pair of outcomes $(Y \le j)$ and $(Y > j)$. The common effect β for each j implies that the three curves have the same shape. As in logistic regression, the size of $|\beta|$ determines how quickly the curves climb or drop. At any fixed x value, the curves have the same ordering as the cumulative probabilities, the one for $P(Y \le 1)$ being lowest.

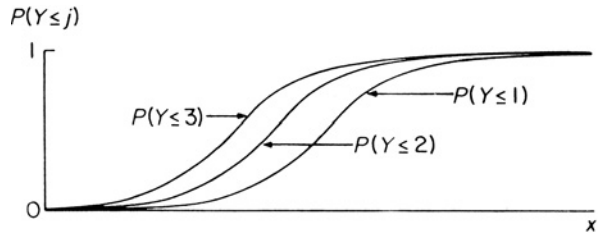


Figure 6.2 Depiction of cumulative probabilities in the cumulative logit model.

Figure 6.2 has $\beta > 0$. Figure 6.3 shows corresponding curves for the category probabilities, $P(Y = j)$. As x increases, $P(Y = 1)$ increases and $P(Y = c)$ decreases.[2] When

[2] Section 6.2.6 shows that a latent variable model implies an ordinal model expressed as $\text{logit}[P(Y \le j)] = \alpha_j - \beta x$; then $\beta > 0$ implies that $P(Y = c)$ increases as x increases.

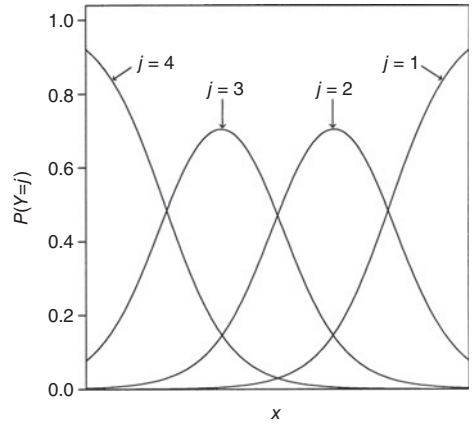


Figure 6.3 Depiction of category probabilities in the cumulative logit model.

$\beta < 0$, the curves in Figure 6.2 descend rather than ascend, and the labels in Figure 6.3 reverse order. When the model holds with $\beta = 0$, the graph has a horizontal line for each cumulative probability. Then x has no effect on Y.

Model interpretations can use odds ratios for the cumulative probabilities and their complements. For two values a and b of x, the *cumulative odds ratio* is

$$\frac{P(Y \le j \mid x = a)/P(Y > j \mid x = a)}{P(Y \le j \mid x = b)/P(Y > j \mid x = b)}.$$

The log of this odds ratio is the difference between the cumulative logits at those two values of x. This equals $\beta(a - b)$, proportional to the distance between the x values. The same proportionality constant (β) applies for each cumulative probability (i.e., for each j). This property is called *proportional odds*. For $a - b = 1$, the property says that the odds of response below any particular category multiply by e^{β} for each 1-unit increase in x.

With multiple explanatory variables, the cumulative logit model with the proportional odds property is

$$\text{logit}[P(Y \le j)] = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \quad j = 1, \ldots, c - 1, \qquad (6.5)$$

with the same effects for each cumulative logit. The ML fitting process uses the Fisher scoring iterative algorithm simultaneously for all j. When we reverse the order of the response categories, the fit is the same but the sign of each $\hat{\beta}_j$ reverses. This sign reversal also applies when we form the log odds by contrasting the high end of the scale with the low end, instead of the low end with the high.

6.2.2 Example: Political Ideology and Political Party Affiliation

Table 6.3, from the 2016 General Social Survey, relates political ideology in the US to political party affiliation. Here, we use only those subjects who identified themselves as "strong

Democrats" or "strong Republicans." Political ideology has a five-point ordinal scale, ranging from very liberal to very conservative. Let x_1 be an indicator variable for political party affiliation (0 = Democrats, 1 = Republicans) and x_2 an indicator variable for gender (0 = females, 1 = males).

Table 6.3 Political ideology by gender and political party affiliation.

		Political Ideology				
Gender	Political Party	Very Liberal	Slightly Liberal	Moderate	Slightly Conservative	Very Conservative
Female	Democrat	25	105	86	28	4
	Republican	0	5	15	83	32
Male	Democrat	20	73	43	20	3
	Republican	0	1	14	72	32

Source: 2016 General Social Survey, `Polviews` data file at text website.

With $c = 5$ response categories, the model has four $\{\alpha_j\}$ intercept terms. Usually, $\{\hat{\alpha}_j\}$ are not of interest except for estimating response probabilities. The estimated effect of political party is $\hat{\beta}_1 = -3.634$ ($SE = 0.218$). For any fixed j, the estimated odds that a Republican's response is in the *liberal* direction rather than the *conservative* direction (i.e., $Y \leq j$ rather than $Y > j$) equal $\exp(\hat{\beta}_1) = \exp(-3.634) = 0.0026$ times the estimated odds for Democrats. The estimated odds that a Republican's response is in the *conservative* direction rather than the *liberal* direction (i.e., $Y > j$ rather than $Y \leq j$) equal $\exp(-\hat{\beta}_1) = \exp(3.634) = 37.9$ times the estimated odds for Democrats. Strong Republicans tend to be much more conservative than strong Democrats. By contrast, the following `R` output does not show evidence of a gender effect:

```
------------------------------------------------------------------------------
> Polviews <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Polviews.dat",
+                        header=TRUE)
> Polviews # grouped data; ungrouped data file at website is Polviews2.dat
  gender party y1  y2 y3 y4 y5
1 female   dem 25 105 86 28  4
2 female repub  0   5 15 83 32
3   male   dem 20  73 43 20  3
4   male repub  0   1 14 72 32
> library(VGAM)
> fit <- vglm(cbind(y1,y2,y3,y4,y5) ~ party + gender,
+             family=cumulative(parallel=TRUE), data=Polviews)
> summary(fit) # "parallel=TRUE" imposes proportional odds structure
                Estimate  Std. Error   z value  Pr(>|z|)
(Intercept):1   -2.12233     0.16875   -12.577    <2e-16 # 4 intercepts for
(Intercept):2    0.16892     0.11481     1.471     0.141 # 5 y categories
(Intercept):3    1.85716     0.15103    12.297    <2e-16
(Intercept):4    4.65005     0.23496    19.791    <2e-16
```

```
partyrepub      -3.63366     0.21785   -16.680    <2e-16 # same efffects
gendermale       0.04731     0.14955     0.316     0.752 # for all 4 logits
---
Residual deviance: 9.8072 on 10 degrees of freedom
--------------------------------------------------------------------------------
```

The model expression for the cumulative probabilities themselves is

$$P(Y \leq j) = \frac{\exp(\alpha_j + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\alpha_j + \beta_1 x_1 + \beta_2 x_2)}.$$

Individual response category probabilities are differences of these, such as

$$P(Y = 3) = P(Y \leq 3) - P(Y \leq 2).$$

Software can display the estimated response category probabilities:

```
--------------------------------------------------------------------------------
> attach(Polviews)
> data.frame(gender, party, fitted(fit)) # y1 = very lib., y5 = very conserv.
  gender   party      y1      y2      y3      y4      y5
1  female     dem  0.1069  0.4352  0.3228  0.1256  0.0095
2  female   repub  0.0031  0.0272  0.1144  0.5895  0.2657
3    male     dem  0.1115  0.4423  0.3165  0.1206  0.0090
4    male   repub  0.0033  0.0284  0.1189  0.5927  0.2566
--------------------------------------------------------------------------------
```

Viewing these helps us to understand the effects of the explanatory variables. For each political party affiliation, the estimated distributions are very similar for females and males. The most common response is *slightly liberal* for Democrats, who are very likely to be in category 3 or below, and *slightly conservative* for Republicans, who are very likely to be in category 4 or 5.

6.2.3 Inference about Cumulative Logit Model Parameters

To check whether political party affiliation has a statistically significant effect, we can conduct a likelihood ratio test of H_0: $\beta_1 = 0$ by comparing the deviance to that of the simpler model without the effect. The simpler model has a residual deviance of 413.05 compared to 9.81 for the full model, so the likelihood-ratio statistic is $413.05 - 9.81 = 403.25$, with $df = 1$. The P-value is essentially 0, extremely strong evidence of an effect. Similar strong evidence results from the Wald test, using $z^2 = (\hat{\beta}_1/SE)^2 = (3.634/0.218)^2 = 278.2$ with $df = 1$. Recall that the likelihood-ratio test is often more powerful than the Wald test for models for categorical responses, especially when the true effects are strong. Here is some R output:

```
--------------------------------------------------------------------------
> fit2 <- vglm(cbind(y1,y2,y3,y4,y5) ~ gender, # removing party effect
+              family=cumulative(parallel=TRUE), data=Polviews)
> lrtest(fit, fit2)
Likelihood ratio test
```

```
Model 1: cbind(y1, y2, y3, y4, y5) ~ party + gender
Model 2: cbind(y1, y2, y3, y4, y5) ~ gender
  #Df    LogLik  Df   Chisq  Pr(>Chisq)
1  10   -35.203
2  11  -236.827   1  403.25   < 2.2e-16
> confint(fit, method="profile")
                 2.5 %    97.5 %
partyrepub    -4.07164  -3.21786  # profile likelihood CI's for
gendermale    -0.24639   0.34140  # beta_1 and beta_2 in full model
-------------------------------------------------------------------------
```

A 95% profile likelihood confidence interval for $-\beta_1$ is (3.207, 4.061). The confidence interval comparing Republicans with Democrats for the cumulative odds ratio with the conservative end of the political ideology scale in the numerator equals $(\exp(3.207), \exp(4.061))$ or (24.7, 58.0). The odds of being in the conservative direction is at least 24.7 times as high for strong Republicans as for strong Democrats. The effect is practically as well as statistically significant.

6.2.4 Increased Power for Ordinal Analyses

For testing independence in contingency tables with ordinal variables, Section 2.5 showed that ordinal test statistics are usually more appropriate and provide greater power than ordinary chi-squared tests that treat the variables as qualitative (nominal-scale). Likewise, cumulative logit models, which utilize the ordinality of Y, usually have a power advantage over baseline-category logit models, which treat Y as nominal-scale.

For instance, for a contingency table with two ordinal variables, a cumulative logit model with proportional odds structure uses a single parameter to describe the association. A baseline-category logit model requires several parameters to do this. Even if the simpler model has some lack of fit, when it captures most of the effect it will tend to produce smaller P-values because of being based on a single degree of freedom.

6.2.5 Example: Happiness and Family Income

Table 6.4, from a General Social Survey, shows the relation between $Y =$ happiness and $x =$ family income. This table has data stratified by race. Here we will analyze the sample of black Americans. As an exercise, you can do a similar analysis for white Americans.

Table 6.4 Happiness and family income for black Americans, with results for white Americans in parentheses.

	Happiness		
Family Income	Not Too Happy	Pretty Happy	Very Happy
Below average	37 (128)	90 (324)	45 (107)
Average	25 (66)	93 (479)	56 (295)
Above average	6 (35)	18 (247)	13 (184)

Source: 2016 General Social Survey, `Happy` data file at text website.

To treat both variables as ordinal, we use a cumulative logit model and regard family income as quantitative with scores (1, 2, 3) for its categories. The likelihood-ratio test of independence compares the model

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta x, \; j = 1, 2,$$

to the null model containing only the intercept terms. The difference between the deviances is 3.11, based on $df = 4 - 3 = 1$. This gives P-value $= 0.078$:

```
-------------------------------------------------------------------------------
> Happy <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Happy.dat",
+                     header=TRUE)
> Happy # data for sampled black Americans
  income y1 y2 y3
1      1 37 90 45
2      2 25 93 56
3      3  6 18 13
> library(VGAM)
> fit <- vglm(cbind(y1,y2,y3)~ income, family=cumulative(parallel=TRUE),
+             data=Happy)
         Estimate  Std. Error  z value  Pr(>|z|)  # not showing the two
income    -0.2668      0.1510   -1.768    0.0771  # intercept estimates
---
> fit0 <- vglm(cbind(y1,y2,y3)~ 1, family=cumulative, data=Happy) # null model
> lrtest(fit, fit0)
Model 1: cbind(y1, y2, y3) ~ income # treating happiness and income as ordinal
Model 2: cbind(y1, y2, y3) ~ 1
  #Df   LogLik  Df  Chisq  Pr(>Chisq)
1   3  -14.566
2   4  -16.121   1  3.109     0.07786 .
-------------------------------------------------------------------------------
```

The estimated family income effect of $\hat{\beta} = -0.267$ suggests that the *not too happy* outcome is less likely as income increases. For the one-sided alternative of an association in the population with this direction, the P-value $= 0.039$.

Baseline-category logit models treat Y as nominal-scale. Such a model also treats x as nominal when we use indicator variables for its categories rather than assume a linear trend for a set of scores for x. For example, let $x_1 = 1$ if family income is in the first category (0 otherwise) and $x_2 = 1$ if family income is in the second category (0 otherwise). The baseline-category logit model with a qualitative income explanatory variable is

$$\log\left(\frac{\pi_j}{\pi_3}\right) = \alpha_j + \beta_{j1}x_1 + \beta_{j2}x_2, \;\; j = 1, 2.$$

The null hypothesis of independence of happiness and income,

$$H_0 : \beta_{j1} = \beta_{j2} = 0, \; j = 1, 2,$$

equates 4 parameters to 0. In this case, the difference of deviances equals 4.13, based on $df = 4$, for which the P-value is 0.39:

```
------------------------------------------------------------------------------
> fit2 <- vglm(cbind(y1,y2,y3) ~ factor(income), family=multinomial,data=Happy)
> fit0 <- vglm(cbind(y1,y2,y3)~ 1, family=multinomial, data=Happy)
> # baseline cat. logit null model equivalent to cumulative logit null model
> lrtest(fit2, fit0)
Model 1: cbind(y1, y2, y3) ~ factor(income) # treats variables as nominal-scale
Model 2: cbind(y1, y2, y3) ~ 1
  #Df   LogLik  Df   Chisq  Pr(>Chisq)
1   0  -14.058                        # fit2 model is saturated
2   4  -16.121   4  4.1258      0.3892
------------------------------------------------------------------------------
```

This test has the advantage of not assuming as much about the model structure, such as linearity for the family income effect. A disadvantage is that it usually has lower power, because the null hypothesis has more parameters. If there truly is a trend in the relationship, we are more likely to capture it with the ordinal analysis, because it focuses the analysis on $df = 1$.

6.2.6 Latent Variable Linear Models Imply Cumulative Link Models

With the proportional odds form of cumulative logit model, the effect of an explanatory variable is the same in the $c - 1$ equations for the different cumulative logits. However, when might such a simple model be plausible? In fact, this proportional odds structure is implied by a simple latent variable model. With many ordinal variables, it is realistic to regard the observed response as a crude measurement of a continuous latent variable. We have modeled political ideology with five categories. With precise measurement of political ideology, we can imagine an essentially continuous response.

Let Y^* denote a latent variable. Let $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_c = \infty$ denote *cutpoints* of the continuous scale for Y^* such that the ordinal categorical variable Y satisfies

$$y = j \quad \text{if } \alpha_{j-1} < y^* \leq \alpha_j.$$

We observe Y in category j when the latent variable falls in the jth interval of values. Figure 6.4 depicts this. Now, suppose the latent variable satisfies an ordinary linear model relating it to the explanatory variables,

$$Y^* = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

where ϵ has some probability distribution with mean 0 and the same variance at all values of the explanatory variables. Then, one can show[3] that the observed ordinal categorical variable satisfies the model

$$\text{link}[P(Y \leq j)] = \alpha_j - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p, \;\; j = 1, \ldots, c-1, \qquad (6.6)$$

for a link function that depends on the distribution of ϵ. With logit link, this is the cumulative logit model (6.5) of proportional odds form, except that the signs of the effects change.

[3] For details, see Agresti (2013), pp. 303–304.

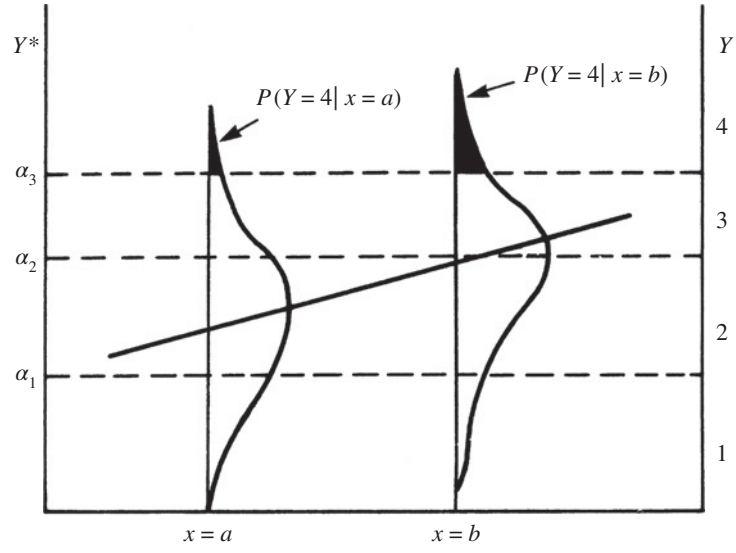


Figure 6.4 Ordinal measurement and underlying regression model for a latent variable. The left vertical axis shows values for the latent variable. The right vertical axis shows values for the observed ordinal response, its category being determined by the three cutpoints on the latent variable scale. The curves show the conditional distribution of the latent variable at two values of the explanatory variable. The line connecting their means represents the regression model for the latent variable.

Because of this latent variable model connection, some software[4] fits the model with this parameterization.

Model (6.6) with its family of possible link functions is called a *cumulative link model*. The link function is such that the shape of the curve for each cumulative probability, when plotted against a quantitative x_j, is the same as the shape of the *cdf* of the distribution of ϵ. Moreover, not only are the implied effects $\{\beta_j\}$ in model (6.6) the same for each cumulative probability, but they are the same as in the latent variable model.

When ϵ has a normal distribution, so that the latent variable model is the ordinary normal linear model, the corresponding link function is the *probit*. Then, a probit model holds for the cumulative probabilities, generalizing the probit model for binary data. (We will discuss the *cumulative probit model* in Section 6.3.6.) When ϵ has a *logistic distribution*, which is bell-shaped and symmetric like the normal but has slightly thicker tails, the corresponding link function is the *logit* and a cumulative logit model holds. The fits with probit and logit links are very similar.

Here is the practical implication of this latent variable connection: if it is plausible to envision that an ordinary linear model with the chosen explanatory variables describes well the effects for an underlying latent variable, then it is sensible to use the cumulative logit model with proportional odds structure or the corresponding cumulative probit model.

6.2.7 Invariance to Choice of Response Categories

In the connection just mentioned between the model for y and a model for a latent variable y^*, the same parameters occur for the effects regardless of how the cutpoints $\{\alpha_j\}$ discretize

[4] Such as the *polr* function in R, the *ologit* function in Stata, and SPSS.

the real line to form the scale for y. The effect parameters are *invariant* to the choice of categories for y.

For example, if a continuous variable measuring political ideology satisfies a linear model with some explanatory variables, then the same effect parameters apply to a discrete version of political ideology with the categories (liberal, moderate, conservative) or (very liberal, slightly liberal, moderate, slightly conservative, very conservative). Therefore, two researchers who use different response categories in studying a predictor's effect should obtain similar $\{\hat{\beta}_j\}$, apart from sampling error. This nice feature of the model makes it possible to compare estimates from studies using different scales for the response variable.

To illustrate, we collapse Table 6.3 to a three-category response, combining the two liberal categories and combining the two conservative categories. Then, the estimated political party effect changes only from -3.634 ($SE = 0.218$) to -3.728 ($SE = 0.229$). Substantive interpretations are unchanged.

Some researchers collapse ordinal responses to binary so they can use ordinary logistic regression. However, a loss of efficiency then occurs, in the sense that larger standard errors result. In practice, when observations are spread fairly evenly among the categories, the efficiency loss is minor when you collapse a large number of categories to 4 or 5 categories, but it can be severe when you collapse to a binary response. It is inadvisable to do this.

6.3 CUMULATIVE LINK MODELS: MODEL CHECKING AND EXTENSIONS *

In this section we present ways of checking the adequacy of cumulative link models. We also present an extension of the cumulative logit model without proportional odds structure, discuss the cumulative link model using the probit link, suggest R^2 and multiple correlation measures of predictive power for the models, and introduce the Bayesian approach for multinomial logit models.

6.3.1 Checking Ordinal Model Goodness of Fit

With discrete explanatory variables and a non-sparse grouped data file for an ordinal model, the residual deviance compares ML fitted counts to the observed counts, such as explained for logistic models in Section 5.2.1 and baseline-category logit models in Section 6.1.4. The deviance is then a test statistic for a global goodness-of-fit test.

For the political ideology data with gender and political party explanatory variables (Table 6.3), the goodness-of-fit test using the deviance is valid, because the grouped data are a non-sparse contingency table. The deviance for the cumulative logit model is 9.81, based on $df = 10$, as shown in the output in Section 6.2.2. The P-value is 0.46, so the model fits adequately. For the more-directed analysis of checking the model by adding an interaction term between gender and political party to the model, the deviance decreases from 9.81 to 8.45, with $df = 1$. This more complex model is not needed (P-value $= 0.24$).

6.3.2 Cumulative Logit Model without Proportional Odds

The cumulative logit model with proportional odds form implies that the distribution of Y at any setting of the explanatory variables is shifted up or shifted down from the distribution of Y at any other setting, or is the same. In the political ideology example, for each gender

Republicans tend to be more conservative than Democrats. When an explanatory variable refers to two groups, as in Table 6.3 for political party or for gender, the model does *not* fit well when the response distributions differ in their *variability*, so that such a shift up or down in the distributions does not occur. If Democrats were primarily moderate in political ideology, while Republicans were very conservative and very liberal, then the variability would be greater for Republicans than Democrats. The two political ideology distributions would be quite different, but the model would not be able to describe this.

When the model fits poorly, we could consider the more general cumulative logit model that has separate effects for the different cumulative probabilities. This model implies that curves for different cumulative probabilities climb or fall at different rates. Therefore, those curves may cross at certain predictor values. This is inappropriate, because this violates the order that cumulative probabilities must have, such as $P(Y \leq 2) \leq P(Y \leq 3)$ for all values of the explanatory variables. However, such a model can fit adequately over a restricted range of explanatory variable values, as often happens with factors as explanatory variables. Sometimes software fails in fitting this more complex model because of violating the order restriction. When we can fit it, an alternative goodness-of-fit test of the model with proportional odds structure is the likelihood-ratio test comparing it to this more complex model by the change in deviance. This test is valid for grouped and ungrouped data files.

For the political ideology data, the more complex model having separate effects for each cumulative probability has deviance 3.59 with $df = 4$, as shown next:

```
------------------------------------------------------------------------------
> library(VGAM)
> summary(vglm(cbind(y1,y2,y3,y4,y5) ~ party + gender, family=cumulative,
+              data=Polviews))         # parallel=FALSE by default
Coefficients: # each cumul. logit has intercept, party effect, gender effect
                Estimate  Std. Error  z value  Pr(>|z|)
(Intercept):1   -2.17521     0.20941  -10.387   < 2e-16
(Intercept):2    0.12173     0.12476    0.976     0.329
(Intercept):3    1.88810     0.17043   11.078   < 2e-16
(Intercept):4    4.10365     0.39770   10.318   < 2e-16
partyrepub:1   -20.76294  3458.04727   -0.006     0.995 # infinite estimate
partyrepub:2    -3.94288     0.42696   -9.235   < 2e-16
partyrepub:3    -3.68095     0.23285  -15.808   < 2e-16
partyrepub:4    -2.94499     0.40785   -7.221  5.17e-13
gendermale:1     0.21835     0.31762    0.687     0.492
gendermale:2     0.18343     0.19352    0.948     0.343
gendermale:3    -0.08638     0.22198   -0.389     0.697
gendermale:4    -0.14633     0.26939   -0.543     0.587
---
Residual deviance: 3.5861 on 4 degrees of freedom
------------------------------------------------------------------------------
```

The likelihood-ratio test that compares this model to the simpler model with proportional odds structure has test statistic $9.81 - 3.59 = 6.22$ with $df = 10 - 4 = 6$. Again, the test suggests that the simpler model is adequate.

Incidentally, the 2×5 marginal table relating party affiliation to political ideology has a 0 count for very liberal Republicans. Because of this, quasi-complete separation occurs

for the model applied solely with the first cumulative logit. The R output above shows an estimated political party effect of -20.76 for the first logit, with a huge standard error, but the actual ML estimate is $-\infty$. Separation results in infinite estimates for multinomial models, much like Section 5.3 explained for binary data.[5]

When the model with a proportional odds structure fits poorly but we cannot fit the non-proportional-odds model, we could instead fit a baseline-category logit model and use the ordinality in an informal way in interpreting the associations. A disadvantage this approach shares with the one just mentioned is the substantial increase in the number of parameters. Even though the model itself may have less bias, estimates of measures of interest such as odds ratios or category probabilities may be poorer because of the lack of model parsimony. Therefore, we do not recommend this approach unless the lack of fit of the ordinal model is severe in a practical sense. Even when the model with proportional odds structure has some lack of fit, it can be useful for overall approximate summaries of effects.

6.3.3 Simpler Interpretations Use Probabilities

To describe an effect, Section 6.2.1 presented an odds ratio interpretation for cumulative logit models. A simpler summary of effects uses category probabilities for Y directly. In the binary case, Section 4.5.1 introduced such interpretations for ordinary logistic regression. For ordinal models, comparisons can use the cumulative probabilities, but it is more informative to focus on probabilities for the extreme (highest and lowest) individual response categories.

To describe effects of a quantitative variable x, we can compare the extreme-category probabilities on Y at maximum and minimum values of x. To describe effects of categorical variables, we compare extreme-category probabilities on Y for different groups. We control for quantitative variables by setting them at their means. We control for qualitative variables by fixing the category, unless there are several, in which case we can set them at the means of their indicator variables.

For binary data, the *average marginal effect* (Section 4.5.2) summarizes the rate of change in $P(Y = 1)$ as a function of each explanatory variable, by finding the average slope of the curves at the observed values. For ordinal main-effects models, the extreme probabilities (but not the others) change in a monotone manner as an explanatory variable increases, so we can report this effect for each of them.

6.3.4 Example: Modeling Mental Impairment

We illustrate probability-based interpretations for modeling the `Mental` data file at the text website,[6] which comes from a study of mental health for a random sample of adult residents of Alachua County, Florida. The study related $Y =$ mental impairment, which is ordinal with categories ($1 =$ well, $2 =$ mild symptom formation, $3 =$ moderate symptom formation, $4 =$ impaired), to two explanatory variables. The life events index x_1 is a numerical composite measure of the number and severity of important life events such as birth of child, new job, divorce, or death in family that occurred to the subject within the past three years.

[5] I. Kosmidis has extended to multinomial models the D. Firth results on bias reduction. See the `brglm2` package in R.

[6] The complete data are shown in Table 6.2 of Agresti (2015).

In this sample of $n = 40$ observations, it varied between 0 and 9, with a mean of 4.3 and standard deviation of 2.7. The socioeconomic status (SES) index x_2 is binary (0 = low, 1 = high). Here, we use the latent variable induced form (6.6) of the cumulative logit model with proportional odds form

$$\text{logit}[P(Y \leq j)] = \alpha_j - \beta_1 x_1 - \beta_2 x_2,$$

because the `polr` (*proportional odds logistic regression*) function in `R` uses it and easily provides estimated probabilities at fixed settings of explanatory variables:

```
-------------------------------------------------------------------------------
> Mental <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Mental.dat",
+                      header=TRUE)
> Mental
   impair ses life
1       1   1    1
...
40      4   0    9
> library(MASS)
> y <- factor(Mental$impair) # polr function requires response to be a factor
> fit <- polr(y ~ life + ses, method="logistic", data=Mental)
> summary(fit) # not showing the 3 intercept parameter estimates
        Value  Std. Error  t value # these are actually z statistics, not t
life   0.3189      0.1210    2.635
ses   -1.1112      0.6109   -1.819
-------------------------------------------------------------------------------
```

The estimate $\hat{\beta}_2 = -1.111$ suggests that mental impairment tends to decrease at the higher level of SES. For the *impaired* outcome, at the mean life events, $\hat{P}(Y = 4) = 0.300$ at low SES and $\hat{P}(Y = 4) = 0.124$ at high SES. The corresponding estimates of $\hat{P}(Y = 1)$ for the *well* outcome are 0.162 at low SES and 0.370 at high SES. Here is how you can obtain such estimated probabilities, after using the `polr` function to fit the model:

```
-------------------------------------------------------------------------
> predict(fit, data.frame(ses=0, life=mean(Mental$life)), type="probs")
     1       2       3       4
0.1618  0.3007  0.2373  0.3002 # predicted outcome prob's, 4 is impaired
> predict(fit, data.frame(ses=1, life=mean(Mental$life)), type="probs")
     1       2       3       4
0.3696  0.3537  0.1530  0.1237
-------------------------------------------------------------------------
```

The estimate $\hat{\beta}_1 = 0.319$ suggests that mental impairment tends to be worse with higher life events. At low SES, $\hat{P}(Y = 4)$ changes from 0.099 to 0.659 as life events increases from its minimum to maximum values; at high SES, it changes from 0.035 to 0.389:

```
-------------------------------------------------------------------------------
> predict(fit, data.frame(ses=0, life=min(Mental$life)), type="probs")
     1       2       3       4 # text describes effects for cat. 4 = impaired
0.4300  0.3408  0.1303  0.0989
```

```
> predict(fit, data.frame(ses=0, life=max(Mental$life)), type="probs")
      1       2       3       4
0.04103  0.1191  0.1805  0.6593
> predict(fit, data.frame(ses=1, life=min(Mental$life)), type="probs")
     1       2       3       4
0.6962  0.2146  0.0543  0.0349
> predict(fit, data.frame(ses=1, life=max(Mental$life)), type="probs")
     1       2       3       4
0.1150  0.2518  0.2440  0.3892
-------------------------------------------------------------------------------
```

Comparing 0.099 to 0.035 at the minimum life events and comparing 0.659 to 0.389 at the maximum life events provides further information about the SES effect. The sample effect is substantial for each explanatory variable.

At the 40 observed values for life events, the rate of change in the estimated probability per unit change in life events averages to -0.057 for the *well* outcome and to 0.048 for the *impaired* outcome. At those 40 observed values, when SES changes from low to high, the estimated probability of the *well* outcome increases by an average of 0.198 and the estimated probability of the *impaired* outcome decreases by an average of 0.171:

```
-------------------------------------------------------------------------------
> ocAME(fit) # ordinal average marginal effect function from
             # www.stat.ufl.edu/~aa/articles/agresti_tarantola_appendix.pdf
$ME.1 # well outcome category
      effect  std.error
ses    0.198      0.104
life  -0.057      0.019
$ME.4 # impaired outcome category
      effect  std.error
ses   -0.171      0.094
life   0.048      0.017
-------------------------------------------------------------------------------
```

6.3.5 A Latent Variable Probability Comparison of Groups

To compare two groups, a useful summary refers to latent variables y_1^* and y_2^* that underlie responses for the two groups. Suppose these are independent, made at any particular setting of the explanatory variables. The summary measure is $P(Y_2^* > Y_1^*)$. When the indicator variable in the model takes value 0 for group 1 and 1 for group 2 and has estimated coefficient $\hat{\beta}$ in a cumulative logit model with proportional odds structure and latent variable parameterization, then we can estimate this probability by[7]

$$\hat{P}(Y_2^* > Y_1^*) = \frac{\exp(\hat{\beta}/\sqrt{2})}{1 + \exp(\hat{\beta}/\sqrt{2})}.$$

[7] For details, see A. Agresti and M. Kateri, *Biometrics*, **73**: 214–219 (2017).

The estimate takes the value $1/2$ when the fitted distributions are identical (i.e., $\hat{\beta} = 0$), and takes values farther from $1/2$ when $\hat{\beta}$ is farther from 0.

We illustrate this with the mental impairment example just presented. We compare the two SES levels by estimating $P(Y_2^* > Y_1^*)$ for underlying latent variables. With $\hat{\beta}_2 = -1.111$, we obtain

$$\hat{P}(Y_2^* > Y_1^*) = \exp(\hat{\beta}_2/\sqrt{2})/[1 + \exp(\hat{\beta}_2/\sqrt{2})] = 0.31.$$

At any particular value for life events, the estimated probability is 0.31 of worse mental impairment at high SES (group 2) than at low SES (group 1).

6.3.6 Cumulative Probit Model

The latent-variable model connection mentioned in Section 6.2.6 implies that a latent linear model with normal response distribution implies a model for cumulative probabilities using the probit link. That model, called a *cumulative probit model*, generalizes the probit model for binary data presented in Section 5.5.3. The effect estimates for the model fit correspond to effect estimates for the latent variable model.

Here is edited R output from fitting the cumulative probit model to the political ideology data[8] of Table 6.3 using the `polr` function, which has effect parameters that are the same as those in the latent variable model when its residual variance is 1:

```
------------------------------------------------------------------------------
> Polviews2 <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                          Polviews2.dat", header=TRUE)
> Polviews2 # ungrouped data file
    subject gender party ideology
1         1 female   dem        1
2         2 female   dem        1
..
661     661   male repub        5
> library(MASS)
> y <- factor(Polviews2$ideology) # polr() requires response to be a factor
> fit.probit <- polr(y ~ party + gender, method="probit", data=Polviews2)
> summary(fit.probit) # not showing the four intercept parameter estimates
                Value  Std. Error   t value
partyrepub    2.03250     0.10996  18.48409
gendermale   -0.00749     0.08562  -0.08747
------------------------------------------------------------------------------
```

The political party estimate of $\hat{\beta}_1 = 2.03$ means that for the normal latent variable model, with higher y^* values representing greater conservatism, the estimated mean political ideology for Republicans is 2.03 higher than the estimated mean for Democrats. This difference is relative to a residual standard deviation of 1.0 for the normal latent response. With an arbitrary standard deviation, we estimate that the two groups have means that differ by 2.03 standard deviations. This is an extremely large effect.

[8] We use the ungrouped data file, which we need for the following section.

6.3.7 R^2 Based on the Latent Variable Model

We can summarize the predictive power for a cumulative link model by approximating R^2 for the latent variable model[9] presented in Section 6.2.6. Let y_i^* denote the value of the latent variable for subject i. The measure for the latent variable model has the usual proportional reduction in variation form

$$R_L^2 = \frac{\sum_i (y_i^* - \bar{y}^*)^2 - \sum_i (y_i^* - \hat{y}_i^*)^2}{\sum_i (y_i^* - \bar{y}^*)^2} = \frac{\sum_i (\hat{y}_i^* - \bar{y}^*)^2}{\sum_i (y_i^* - \bar{y}^*)^2},$$

where the subscript L reminds us that this is for the latent variable and is unobserved. This equals the estimated variance of $\hat{y}^*$ divided by the estimated variance of y^*. After fitting a model to the ungrouped data file, we can estimate the variance of $\hat{y}^*$ by the variance of the linear predictor, without the intercept terms. We cannot observe the latent variable or its sample variance, but we can approximate that variance by the estimated variance of $\hat{y}^*$ plus the variance of the residual error in the latent variable model. That variance is 1.0 for the standard normal distribution and $\pi^2/3 = 3.29$ for the standard logistic distribution that yield the probit and logit link functions. We then divide the estimated variance of $\hat{y}^*$ by the approximate variance of y^* to obtain the R^2 approximation for R_L^2. Its positive square root R approximates the multiple correlation R_L for the latent variable model.

To illustrate, we find R^2 and the multiple correlation for the cumulative logit model for political ideology, using party affiliation and gender as the explanatory variables. The following R output[10] continues the analysis from the output above using the ungrouped data file:

```
------------------------------------------------------------------------------
> fit.logit <- polr(y ~ party + gender, method="logistic", data=Polviews2)
> var(fit.logit$lp)/(var(fit.logit$lp) + (pi^2)/3) # lp = linear predictor
[1] 0.48698 # R-squared based on logistic latent var. model for cumul. logit
> sqrt(0.48698)
[1] 0.69784 # multiple correlation based on logistic latent variable model
> var(fit.probit$lp)/(var(fit.probit$lp) + 1.0)
[1] 0.4945 # R-squared based on normal latent var. model for cumulative probit
------------------------------------------------------------------------------
```

We predict that 48.7% of the variability in the political ideology latent variable is explained by the two explanatory variables. This moderately large value reflects that the data set used only the extreme categories of party affiliation (i.e., *strong* Democrats and *strong* Republicans), and those two categories predict political ideology fairly well. In fact, we obtain $R^2 = 0.487$ with party affiliation as the sole explanatory variable in the model. When we use both explanatory variables and add an interaction term, R^2 increases only to 0.488. We get similar results with the cumulative probit model, using the fit from the output above and using 1.0 for the normal residual variance.

[9] Some software (such as Stata) calls this measure the *McKelvey–Zavoina* R^2, named after the authors of a 1975 article that proposed it for cumulative probit models.

[10] The measure will be available with function `R2latvar` in the next release of VGAM.

6.3.8 Bayesian Inference for Multinomial Models

For Bayesian inference for multinomial models, as with Bayesian inference for binary regression models (Section 5.4), it is common to use diffuse normal prior distributions for the effect parameters. For any link function, Bayesian model fitting uses MCMC with the product of the chosen prior densities and the multinomial likelihood function for the model.

For cumulative link models, prior distributions for the intercept parameters $(\alpha_1, \ldots, \alpha_{c-1})$ should take into account the ordering constraint

$$-\infty < \alpha_1 < \alpha_2 < \cdots < \alpha_{c-1} < \infty,$$

such as by appropriately truncating priors that would be used without the constraint. For the baseline-category logit model (6.3), if you place simple structure such as a common variance for the priors for parameters $\beta_{1k}, \beta_{2k}, \ldots, \beta_{c-1,k}$ for the effects for explanatory variable k, posterior results then depend slightly on the choice of baseline category, because an effect relative to a pair of nonbaseline categories, $\beta_{jk} - \beta_{j'k}$, then has twice the prior variance. Alternatively, you can parameterize the model so that each logit has the same prior variance. The same remark applies to factors in such models; results should be invariant to the choice of a baseline category for indicators. For a binary factor, for example, we can use indicator values 0.5 and -0.5 instead of 1 and 0, so that each category of the factor has a parameter with the same variance.

6.3.9 Example: Modeling Mental Impairment Revisited

To illustrate the Bayesian approach, we show results[11] for an analysis of the mental impairment data file for which Section 6.3.4 showed a frequentist analysis. We use relatively flat normal priors for the parameters, with means of 0 and standard deviations of 10. The posterior mean estimates reported in the next R output, based on a long MCMC process, refer to the latent variable parameterization of the model, for which log cumulative odds ratios contrast the high with the low end of the response scale:

```
-------------------------------------------------------------------------------
> Mental2 # data file at text website
   impair   ses  life
1    imp1   0.5     1
2    imp1   0.5     9
...
40   imp4  -0.5     9
> fit_freq <- polr(impair ~ life + ses, method="logistic", data=Mental2)
> summary(fit_freq) # frequentist analysis (Section 6.3.4)
        Value  Std. Error  t value
life   0.3189      0.1210    2.635
ses   -1.1112      0.6109   -1.819

> library(BayesOrd) # obtain from https://github.com/tjmckinley/BayesOrd
> # e.g., using install_github("tjmckinley/BayesOrd") with devtools package
```

[11] Thanks to T.J. McKinley for help with his R package BayesOrd, developed for analyses described in *Bayesian Analysis* **10**: 1–30, by him, M. Morters, and J.L.N. Wood.

```
> fit <- bayesord(impair ~ life + ses, fixed=TRUE, mnb=0, varb=100, vart=100,
+        niter=1e+06, nchains=2, start=10000, data=Mental2)
>        # mnb, varb are mean, var. of beta's; vart =  var. of intercepts
> summary(fit, digits=3) # for parameterization alpha_j - beta1 x1 - beta2 x2
              Mean     SD  Median    2.5%   97.5%  MC error/SD
logOR(life)  0.357  0.126   0.353   0.118  0.6130       0.00743
logOR(ses)  -1.210  0.634  -1.200  -2.470  0.0146       0.00207
> props <- as.matrix(fit$beta)
> apply(props, 2, function(x)sum(x > 0)/length(x))
   life      ses
0.99874  0.02646 # posterior P(beta > 0) for each effect parameter
-------------------------------------------------------------------------------
```

The estimated cumulative log odds ratios are similar in magnitude to those in the frequentist analysis, but have Bayesian interpretations. For example, the posterior probability is 0.95 that the odds that mental impairment at high SES is *impaired* (instead of better) are between $\exp(-2.470) = 0.08$ and $\exp(0.0146) = 1.01$ times the corresponding odds at low SES. Corresponding to the frequentist P-value for testing H_0: $\beta_2 = 0$ against H_a: $\beta_2 < 0$ is the Bayesian posterior probability that $\beta_2 \geq 0$. This is 0.026, compared to $0.070/2 = 0.035$ for the one-sided P-value from the Wald test.

For further details about Bayesian analyses for multinomial models, see Agresti (2013, Section 8.6) and Hoff (2009, Chapter 12).

6.4 PAIRED-CATEGORY LOGIT MODELING OF ORDINAL RESPONSES*

Cumulative logit models for ordinal responses use the entire response scale in forming each logit. This section introduces logit models for ordered categories that, like baseline-category logit models, use *pairs* of categories. Odds ratio interpretations can then use individual categories instead of cumulative probabilities and their complements. Again, models that assume proportional odds structure have fewer parameters.

6.4.1 Adjacent-Categories Logits

An alternative ordinal logit model forms logits for all pairs of adjacent categories. The *adjacent-categories logits* are

$$\log\left(\frac{\pi_j}{\pi_{j+1}}\right), \quad j = 1, \ldots, c-1.$$

For $c = 3$, these logits are $\log(\pi_1/\pi_2)$ and $\log(\pi_2/\pi_3)$. With an explanatory variable x, an adjacent-categories logit model with proportional odds structure has the form

$$\log\left(\frac{\pi_j}{\pi_{j+1}}\right) = \alpha_j + \beta x, \quad j = 1, \ldots, c-1. \tag{6.7}$$

For it, the effect β of x on the odds of making the lower instead of the higher response is identical for each pair of adjacent response categories. Like the cumulative logit model (6.4)

of proportional odds form, this model has a single parameter rather than $c-1$ parameters for the effect of x. This provides advantages of parsimony, such as a simpler summary of the effect.

The adjacent-categories logits, like the baseline-category logits, determine the logits for all pairs of response categories. For model (6.7), the coefficient of x for $\log(\pi_a/\pi_b)$ equals $\beta(b-a)$. The effect depends on the distance between categories, so this model utilizes the ordering of the response scale.

6.4.2 Example: Political Ideology Revisited

We return to Table 6.3 and model political ideology in terms of $x_1 =$ party affiliation and $x_2 =$ gender, now using the adjacent-categories logit model,

$$\log\left(\frac{\pi_j}{\pi_{j+1}}\right) = \alpha_j + \beta_1 x_1 + \beta_2 x_2, \;\; j = 1, 2, 3, 4.$$

Here is edited R output for the model fit:

```
------------------------------------------------------------------------------
> library(VGAM)
> fit <- vglm(cbind(y1,y2,y3,y4,y5) ~ party + gender,
+             family=acat(parallel=TRUE, reverse=TRUE), data=Polviews)
> summary(fit) # family=acat gives adjacent-category logits
             Estimate  Std. Error  z value  Pr(>|z|) # not showing the 4
partyrepub  -2.23478     0.16841  -13.270   < 2e-16 # intercept est's
gendermale   0.01212     0.09661    0.125   0.90016
---
Names of linear predictors: # reverse=FALSE would use log(P[Y=j+1]/P[Y=j])
loge(P[Y=1]/P[Y=2]),loge(P[Y=2]/P[Y=3]),loge(P[Y=3]/P[Y=4]),loge(P[Y=4]/P[Y=5])
Residual deviance: 13.4665 on 10 degrees of freedom
------------------------------------------------------------------------------
```

The party affiliation effect is $\hat{\beta}_1 = -2.235$. The estimated odds that a Democrat's political ideology is in category j instead of $j+1$ are[12] $\exp(-\hat{\beta}_1) = 9.34$ times the estimated odds for Republicans. For each gender, this is the estimated odds ratio for each of the four 2×2 tables consisting of a pair of adjacent columns of Table 6.3. This type of odds ratio is called a *local odds ratio*, because it refers to the effect for a localized part of the response scale. The estimated odds ratio for an arbitrary pair of columns $a < b$ equals $\exp[\hat{\beta}_1(b-a)]$. For example, the estimated odds that a Democrat's ideology is *very liberal* (category 1) instead of *very conservative* (category 5) are $\exp[2.235(5-1)] = (9.34)^4 = 7624$ times those for Republicans!

The model fit has deviance $G^2 = 13.47$ with $df = 10$, an adequate fit. Substantive inferential results about effects are similar to those for the cumulative-logit analysis in Section 6.2.2.

[12] We take $-\hat{\beta}_1$ in the exponent because Democrat is value 0 for the party indicator.

6.4.3 Sequential Logits

Another approach forms logits for ordered response categories in a sequential manner. The models apply simultaneously to

$$\log\left(\frac{\pi_1}{\pi_2+\cdots+\pi_c}\right), \log\left(\frac{\pi_2}{\pi_3+\cdots+\pi_c}\right), \ldots, \log\left(\frac{\pi_{c-1}}{\pi_c}\right).$$

For $c = 3$, these logits are $\log[\pi_1/(\pi_2+\pi_3)]$ and $\log(\pi_2/\pi_3)$. These are called *sequential logits*, sometimes also called *continuation-ratio logits*. They refer to a binary response that contrasts each category with a grouping of all categories from *higher* levels of the response scale.

The sequential logit model form is useful when a sequential mechanism, such as survival through various age periods, determines the response outcome. Let $\omega_j = P(Y = j \mid Y \geq j)$. The sequential logits are ordinary logits of these conditional probabilities: namely, $\log[\omega_j/(1-\omega_j)]$.

6.4.4 Example: Tonsil Size and Streptococcus

We illustrate sequential logits using a study[13] that cross-classified a sample of children by their tonsil size and by whether they were carriers of *Streptococcus pyogenes*, a bacteria that is the cause of Group A streptococcal infections. The tonsil size response variable has three ordered outcomes: (not enlarged, enlarged, greatly enlarged). The counts in those categories were (19, 29, 24) for carriers and (497, 60, 269) for noncarriers.

Sequential logits are natural for these data, because of the sequential process by which a subject can develop greatly enlarged tonsils. The tonsils start in the not-enlarged state and may become enlarged, perhaps explained by some explanatory variable. If the process continues, the tonsils may become greatly enlarged. We use sequential logits to model (1) the probability π_1 of nonenlarged tonsils and (2) the conditional probability $\pi_2/(\pi_2+\pi_3)$ of enlarged tonsils, given that the tonsils were enlarged or greatly enlarged.

Let x indicate whether a child is a carrier of *Streptococcus pyogenes* ($1 =$ yes, $0 =$ no). The sequential logit model with proportional odds structure is

$$\log\left(\frac{\pi_1}{\pi_2+\pi_3}\right) = \alpha_1 + \beta x, \quad \log\left(\frac{\pi_2}{\pi_3}\right) = \alpha_2 + \beta x.$$

The ML estimate of the carrier effect is $\hat{\beta} = -0.528$ ($SE = 0.198$), for which $\exp(\hat{\beta}) = 0.59$. For instance, given that the tonsils were enlarged, the estimated odds for carriers of having enlarged rather than greatly enlarged tonsils were 0.59 times the estimated odds for non-carriers. The model fits the data very well, with deviance 0.006 ($df = 1$). Edited R output follows:

```
-------------------------------------------------------------------------
> Tonsils
  carrier  y1  y2  y3
1     yes  19  29  24
2      no 497 560 269
```

[13] M. Holmes and R. Williams, *J. Hyg. Camb.* **52**: 165–179 (1954).

```
> library(VGAM)
> fit <- vglm(cbind(y1,y2,y3) ~ carrier, family=sratio(parallel=TRUE),
+             data=Tonsils) # family=sratio gives sequential ratio logits
> summary(fit)
                Estimate  Std. Error  z value  Pr(>|z|)
(Intercept):1   -0.51102     0.05614   -9.102   < 2e-16
(Intercept):2    0.73218     0.07286   10.049   < 2e-16
carrieryes      -0.52846     0.19775   -2.672   0.00753
---
Names of linear predictors: logit(P[Y=1|Y>=1]), logit(P[Y=2|Y>=2])
Residual deviance: 0.0057 on 1 degrees of freedom
------------------------------------------------------------------------------
```

For this model, $\exp(\hat{\beta}) = 0.59$ estimates an assumed common value for a cumulative odds ratio from the first part of the model and a local odds ratio from the second part of the model. By contrast, from analyses not shown here, the cumulative logit model of proportional odds form estimates a common value of $\exp(-0.603) = 0.55$ for each cumulative odds ratio (deviance $= 0.30$, $df = 1$), and the adjacent-categories logit model of proportional odds form estimates a common value of $\exp(-0.429) = 0.65$ for each local odds ratio (deviance $= 0.24$, $df = 1$). According to the deviance, any of these three models is plausible. The data provide strong evidence of an association (e.g., Wald P-value $= 0.008$).

EXERCISES

6.1 A model fit predicting preference for President in the US (Democrat, Republican, Independent) using $x =$ annual income (in \$10,000 dollars) is $\log(\hat{\pi}_D/\hat{\pi}_I) = 3.3 - 0.2x$ and $\log(\hat{\pi}_R/\hat{\pi}_I) = 1.0 + 0.3x$.

a. State the prediction equation for $\log(\hat{\pi}_R/\hat{\pi}_D)$. Interpret its slope.

b. Find the range of x for which $\hat{\pi}_R > \hat{\pi}_D$.

c. State the prediction equation for $\hat{\pi}_I$.

6.2 For the alligator food choice example (Section 6.1.2), use the model fit to estimate an odds ratio that describes the effect of length on primary food choice being either *invertebrate* or *other*.

6.3 Table 6.5, available in the `Alligators2` data file at the text website, displays primary food choice for a sample of alligators, classified by length ($\leq$2.3 meters, $>$2.3 meters) and by the lake in Florida in which they were caught. Fit a model to describe the effects of length and lake on primary food choice. Report the prediction equations, with *fish* as the baseline category. Interpret the effect of length on the choice between invertebrates and fish.

6.4 For the belief in an afterlife example (Section 6.1.5), describe the gender effect by reporting and interpreting the estimated conditional odds ratio for the (**a**) *undecided* and *no* pair of response categories, (**b**) *yes* and *undecided* pair. Interpret.

Table 6.5 Data on alligator food choice for Exercise 6.3.

		Primary Food Choice				
Lake	Length	Fish	Invertebrate	Reptile	Bird	Other
Hancock	≤2.3	23	4	2	2	8
	>2.3	7	0	1	3	5
Oklawaha	≤2.3	5	11	1	0	3
	>2.3	13	8	6	1	0
Trafford	≤2.3	5	11	2	1	5
	>2.3	8	7	6	3	5
George	≤2.3	16	19	1	2	3
	>2.3	17	1	0	1	3

Source: Wildlife Research Laboratory, Florida Game and Fresh Water Fish Commission.

6.5 For a recent General Social Survey, a prediction equation relating $Y =$ job satisfaction (4 ordered categories; 1 = least satisfied) to the subject's report of $x_1 =$ earnings compared to others with similar positions (4 ordered categories; 1 = much less, 4 = much more), $x_2 =$ freedom to make decisions about how to do job (4 ordered categories; 1 = very true, 4 = not at all true), and $x_3 =$ work environment allows productivity (4 ordered categories; 1 = strongly agree, 4 = strongly disagree), was $\text{logit}[\hat{P}(Y \leq j)] = \hat{\alpha}_j - 0.54x_1 + 0.60x_2 + 1.19x_3$.

a. Summarize each partial effect by indicating whether subjects tend to be more satisfied or less satisfied, as (i) x_1, (ii) x_2, (iii) x_3, increases.

b. Report the settings for x_1, x_2, x_3 at which a subject is most likely to have the highest job satisfaction.

6.6 Is marital happiness associated with family income? For a General Social Survey, counts in the happiness categories (not, pretty, very) were (6, 43, 75) for a below average income, (6, 113, 178) for an average income, and (6, 57, 117) for an above average income. Table 6.6 shows the output for a baseline-category logit model with

Table 6.6 Software output on modeling happiness for Exercise 6.6.

```
-------------------------------------------------------------------------------
               Estimate  Std. Error  z value  Pr(>|z|)
(Intercept):1  -2.55518     0.72560   -3.521  0.000429
(Intercept):2  -0.35129     0.26837   -1.309  0.190554
income:1       -0.22751     0.34120   -0.667  0.504907
income:2       -0.09615     0.12202   -0.788  0.430694
---
Residual deviance: 3.1909 on 2 degrees of freedom
> fitted(fit)
       y1       y2       y3
1 0.03637  0.37579  0.58784
2 0.03024  0.35625  0.61352
3 0.02506  0.33665  0.63829
> deviance(vglm(cbind(y1,y2,y3) ~ 1, family=multinomial))
[1] 4.13476
-------------------------------------------------------------------------------
```

very happy as the baseline category and scores (1, 2, 3) for the income categories. Prepare a short report, summarizing what you learn from this output.

6.7 Refer to the previous exercise. Table 6.7 shows output for a cumulative logit model with scores (1, 2, 3) for the income categories.

a. Explain advantages of this model over the baseline-category logit model of the previous exercise.

b. Does the model fit adequately? Justify your answer. Why does the output report two intercepts but one income effect?

c. Report a test statistic and P-value for testing that marital happiness is independent of family income. Interpret the income effect.

Table 6.7 Software output on modeling happiness for Exercise 6.7.

```
-------------------------------------------------------------------------------
               Estimate  Std. Error  z value  Pr(>|z|)
(Intercept):1   -3.2466      0.3404   -9.537    <2e-16
(Intercept):2   -0.2378      0.2592   -0.917     0.359
income          -0.1117      0.1179   -0.948     0.343
---
Residual deviance: 3.2472 on 3 degrees of freedom
-------------------------------------------------------------------------------
```

6.8 Table 6.8 results from a clinical trial for the treatment of small-cell lung cancer. Patients were randomly assigned to two treatment groups. The sequential therapy administered the same combination of chemotherapeutic agents in each treatment cycle. The alternating therapy used three different combinations, alternating from cycle to cycle. Fit a cumulative logit model with a proportional odds structure. Interpret the estimated treatment effect. Check whether a model allowing interaction provides a significantly better fit.

Table 6.8 Data for Exercise 6.8 on lung cancer treatment.

			Response to Chemotherapy		
Treatment Therapy	Gender	Progressive Disease	No Change	Partial Remission	Complete Remission
Sequential	Male	28	45	29	26
	Female	4	12	5	2
Alternating	Male	41	44	20	20
	Female	12	7	3	1

Source: Holtbrugge, W., and Schumacher, M., *Appl. Statist.* **40**: 249–259 (1991).

6.9 A cumulative logit model is fitted to data from a General Social Survey, with $Y =$ political ideology (very liberal to very conservative) and $x =$ religious preference (Protestant, Catholic, Jewish, Other). With indicator variables for the first three religion categories, the ML fit has $\hat{\alpha}_1 = -1.03$, $\hat{\alpha}_2 = -0.13$, $\hat{\alpha}_3 = 1.57$, $\hat{\alpha}_4 = 2.41$, $\hat{\beta}_1 = -1.27$, $\hat{\beta}_2 = -1.22$, $\hat{\beta}_3 = -0.44$.

a. How many categories does Y have? Which group is estimated to be the (i) most liberal, (ii) most conservative?

b. Use an estimated odds ratio to compare political ideology for the Protestant and Catholic groups. Interpret.

6.10 For the mental impairment example (Section 6.3.4), when we add an interaction term, we obtain the fit

$$\text{logit}[\hat{P}(Y \leq j)] = \hat{\alpha}_j - 0.420x_1 + 0.371x_2 + 0.181x_1x_2.$$

The coefficient 0.181 of x_1x_2 has $SE = 0.238$. Find the estimated effect of life events for the low SES group ($x_2 = 0$) and for the high SES group ($x_2 = 1$). Explain why these suggest that the impact of life events may be more severe for the low SES group. Does the difference in effects seem to be significant?

6.11 Refer to the previous exercise and the `Mental` data file.

a. For the main effects fit, summarize the life events effect with probability interpretations for $P(Y = 1)$ analogous to those given in the text for $P(Y = 4)$.

b. Find and interpret the estimate of R^2 and the multiple correlation for the corresponding latent variable model.

c. Compare the estimated effects to those obtained after combining response categories 3 and 4. What property of the model does this reflect?

d. Since $n = 40$ is not large, the choice of prior distribution in a Bayesian analysis can be influential. Illustrate this by repeating the analysis of Section 6.3.9 using $\sigma = 1$ in the normal priors. Compare to the results with $\sigma = 10$.

6.12 Refer to the previous two exercises. Fit the main-effects cumulative probit model. Interpret the SES effect. Find and interpret the estimated R^2 and multiple correlation for the corresponding latent variable model.

6.13 Using the output in Section 6.2.2 for the cumulative logit model fitted to the political ideology data, find $\hat{P}(Y_2^* > Y_1^*)$ for comparing strong Republicans with strong Democrats on the underlying latent variable, adjusting for gender. Interpret.

6.14 Table 6.9 shows cell counts from the 2010 General Social Survey that compare Republicans and Democrats when asked whether claims about environmental threats are exaggerated. Fit a multinomial model using frequentist or Bayesian methods. Summarize your analyses in a short report, including edited output, in an appendix.

Table 6.9 Software output on environmental threats for Exercise 6.9.

```
-------------------------------------------------------------------------------
Rows: party    Columns: Environmental threats exaggerated
                        agree   neutral   disagree
Republican                172        57         82
Democrat                  111        78        283
-------------------------------------------------------------------------------
```

6.15 For Table 6.2 on belief in an afterlife, fit a model using (**a**) cumulative logits, (**b**) adjacent-categories logits, and (**c**) sequential logits. Prepare a short report, summarizing your analyses and interpreting and comparing results.

6.16 Table 6.10 refers to a study that randomly assigned subjects to a control group or a treatment group. Daily during the study, treatment subjects ate cereal containing psyllium. The purpose of the study was to analyze whether this resulted in lowering LDL cholesterol. Using a frequentist or Bayesian approach, model the ending cholesterol level as a function of treatment, treating the beginning level (**a**) as a covariate with sensible scores and (**b**) as a categorical control variable. In each case, analyze the treatment effect. Compare results and interpret.

Table 6.10 Data for Exercise 6.16 on cholesterol study.

	Ending LDL Cholesterol Level							
	Control				Treatment			
Beginning	$\leq$3.4	3.4–4.1	4.1–4.9	>4.9	$\leq$3.4	3.4–4.1	4.1–4.9	>4.9
$\leq$3.4	18	8	0	0	21	4	2	0
3.4–4.1	16	30	13	2	17	25	6	0
4.1–4.9	0	14	28	7	11	35	36	6
>4.9	0	2	15	22	1	5	14	12

Source: Dr. Sallee Anderson, Kellogg Co.

6.17 The output in Table 6.11 shows sequential-logit modeling of data from a developmental toxicity study.[14] Rodent studies are commonly used to test and regulate substances posing a potential danger to developing fetuses. This study administered diethylene

Table 6.11 Software output for the toxicity study of Exercise 6.17.

```
-------------------------------------------------------------------------------
> Toxicity # data file at text website
  concentration dead malformation normal
            0.0   15            1     281
2          62.5   17            0     225
3         125.0   22            7     283
4         250.0   38           59     202
5         500.0  144          132       9
> fit <- vglm(cbind(dead,malformation,normal) ~ concentration, family=sratio,
+             data=Toxicity) # using function in VGAM library
> summary(fit)
                Estimate  Std. Error  z value  Pr(>|z|)
(Intercept):1   -3.24793     0.15766   -20.60    <2e-16
(Intercept):2   -5.70190     0.33065   -17.24    <2e-16
concentration:1  0.00639     0.00043    14.70    <2e-16
concentration:2  0.01737     0.00121    14.33    <2e-16
---
Names of linear predictors: logit(P[Y=1|Y>=1]), logit(P[Y=2|Y>=2])
Residual deviance: 11.8384 on 6 degrees of freedom
-------------------------------------------------------------------------------
```

[14] Based on results in C.J. Price et al., *Fund. Appl. Toxicol.* **8**: 115–126 (1987). Thanks to Dr. Louise Ryan for these data, which are in the `Toxicity` data file at the text website.

glycol dimethyl ether, an industrial solvent used in the manufacture of protective coatings, to pregnant mice. Each mouse was exposed to one of five concentration levels for ten days early in the pregnancy. Two days later, the uterine contents of the pregnant mice were examined for defects. Each fetus had the three possible outcomes (dead, malformation, normal). The outcomes are ordered. Prepare a short report that summarizes information from these analyses.

6.18 A response scale has the categories (strongly agree, mildly agree, mildly disagree, strongly disagree, don't know). Describe a way to model this response variable simultaneously using two models.

6.19 Table 6.12, from the `Accidents` data file at the text website, is an expanded version of data that Section 7.2.7 analyzes about a sample of auto accidents. The response categories are (1) not injured, (2) injured but not transported by emergency medical services, (3) injured and transported by emergency medical services but not hospitalized, (4) injured and hospitalized but did not die, (5) injured and died. Analyze these data. Prepare a short report, summarizing your analyses.

Table 6.12 Data for Exercise 6.19 on auto accidents.

			Severity of Injury				
Gender	Location	Seat-Belt	1	2	3	4	5
Female	Urban	No	7287	175	720	91	10
		Yes	11587	126	577	48	8
	Rural	No	3246	73	710	159	31
		Yes	6134	94	564	82	17
Male	Urban	No	10381	136	566	96	14
		Yes	10969	83	259	37	1
	Rural	No	6123	141	710	188	45
		Yes	6693	74	353	74	12

Source: Dr. Cristanna Cook, Medical Care Development, Augusta, Maine.

6.20 True, or false?

a. One reason it is usually wise to treat an ordinal variable with methods that use the ordering is that in tests about effects, chi-squared statistics have smaller df values and tend to be more powerful.

b. The cumulative logit model assumes that Y is ordinal; it should not be used with nominal Y. The baseline-category logit model treats Y as nominal; it can be used with ordinal Y, but it then ignores the ordering.

c. If political ideology is mainly in the moderate category in New Zealand and mainly in the liberal and conservative categories in Australia, models with the proportional odds assumption should fit well for comparing these countries.

CHAPTER 7

LOGLINEAR MODELS FOR CONTINGENCY TABLES AND COUNTS

Loglinear models are generalized linear models (GLMs) for *count data* (Section 3.3). The model expresses counts or rates in terms of explanatory variables that can be categorical and quantitative. One of their main uses is modeling cell counts in contingency tables that cross-classify categorical response variables. Parameters in the model describe associations among the categorical variables.

After introducing loglinear models for contingency tables, we present statistical inference for model parameters and methods for model checking. When one variable is a binary response variable, logistic regression models for that response are equivalent to certain loglinear models. However, loglinear models are mainly useful when at least two variables in a contingency table are response variables. Graphical representations can portray a loglinear model's conditional independence and association patterns.

Ordinary loglinear models for contingency tables treat all variables as nominal scale, but we also present a loglinear model for association between ordinal variables. We also present loglinear models that apply with a single response variable that is a count or a rate. Such models commonly assume that the counts are independent *Poisson* variates, but we also introduce a distribution (the *negative binomial*) that can handle overdispersion — more variability than the Poisson allows.

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

7.1 LOGLINEAR MODELS FOR COUNTS IN CONTINGENCY TABLES

To begin, consider an $r \times c$ contingency table that cross-classifies n subjects on two categorical response variables, a row variable X and a column variable Y. When X and Y are statistically independent, the joint cell probabilities $\{\pi_{ij} = P(X = i, Y = j)\}$ are determined by the row and column marginal probabilities,

$$\pi_{ij} = P(X = i)P(Y = j) = \pi_{i+}\pi_{+j},\ i = 1, \ldots, r,\ j = 1, \ldots, c.$$

The cell probabilities $\{\pi_{ij}\}$ are parameters for a *multinomial* distribution. Loglinear model formulas use expected frequencies $\{\mu_{ij} = n\pi_{ij}\}$ rather than $\{\pi_{ij}\}$. Then they apply also to the *Poisson* distribution for cell counts with expected values $\{\mu_{ij}\}$. Under independence, $\mu_{ij} = n\pi_{i+}\pi_{+j}$ for all i and j.

7.1.1 Loglinear Model of Independence for Two-Way Contingency Tables

The independence condition, $\mu_{ij} = n\pi_{i+}\pi_{+j}$, is multiplicative. Taking the log of both sides of the equation yields an additive relation. That is, independence has the form[1]

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y, \tag{7.1}$$

with an intercept term based on the sample size, a *row effect* term λ_i^X based on the probability in row i, and a *column effect* term λ_j^Y based on the probability in column j. This model is called the *loglinear model of independence*. The larger the value of λ_i^X, the larger each expected frequency is in row i. The larger the value of λ_j^Y, the larger each expected frequency is in column j.

The null hypothesis of independence is, equivalently, the null hypothesis that this loglinear model holds. The fitted values that satisfy the model, $\{\hat{\mu}_{ij} = n_{i+}n_{+j}/n\}$, are the estimated expected frequencies for chi-squared tests of independence (Section 2.4). Those significance tests are also goodness-of-fit tests of this loglinear model.

Loglinear models for contingency tables are generalized linear models that treat the cell counts as independent observations from Poisson distributions and use the log link function. As formula (7.1) illustrates, loglinear models do not distinguish between response and explanatory classification variables. Model (7.1) specifies how the expected cell counts vary according to the categories of X and Y. The model regards the observations to be the cell counts rather than the classifications of individual subjects.

7.1.2 Interpretation of Parameters in the Independence Model

Parameter interpretation is simplest when we view one response variable as a function of the other. For $r \times 2$ contingency tables, for instance, the logit in row i is

$$\begin{aligned}\text{logit}[P(Y = 1)] &= \log\left[\frac{P(Y = 1)}{1 - P(Y = 1)}\right] = \log\left(\frac{\mu_{i1}}{\mu_{i2}}\right) = \log \mu_{i1} - \log \mu_{i2} \\ &= (\lambda + \lambda_i^X + \lambda_1^Y) - (\lambda + \lambda_i^X + \lambda_2^Y) = \lambda_1^Y - \lambda_2^Y.\end{aligned}$$

[1] The X and Y superscripts are labels for the variables, not power exponents.

This logit for Y does not depend on the level i of X. In each row, the odds of response in column 1 equal $\exp(\lambda_1^Y - \lambda_2^Y)$. In model (7.1), differences between two parameters for a variable relate to the log odds of making one response, relative to another, at each category for the other variable.

The independence model formula treats X and Y as factors, so one of $\{\lambda_i^X\}$ is redundant and one of $\{\lambda_j^Y\}$ is redundant, because parameterization requires one fewer indicator variable than the number of factor levels. Software sets the parameter equal to 0 for either the first category (as in R) or the last category (as in SAS). What *is* unique is the *difference* between two main effect parameters of a particular type. That is what determines odds and odds ratios.

7.1.3 Example: Happiness and Belief in Heaven

In a recent General Social Survey, subjects in the US were asked about their $X =$ happiness (not too happy, pretty happy, very happy) and about $Y =$ whether they believed in heaven (no, yes). The following R output shows results of fitting the independence loglinear model to the 3×2 contingency table, but with the data file specified in terms of the six modeled cell counts:

```
------------------------------------------------------------------------------
> HappyHeaven <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                           HappyHeaven.dat", header=TRUE)
> HappyHeaven    # Data file HappyHeaven at text website
    happy   heaven  count
1     not       no     32
2     not      yes    190
3  pretty       no    113
4  pretty      yes    611
5    very       no     51
6    very      yes    326
> fit <- glm(count ~ happy + heaven, family=poisson, data=HappyHeaven)
> # canonical link for Poisson is log, so "(link=log)" is not necessary
> # loglm function in MASS library also fits loglinear models
> summary(fit) # independence loglinear model
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   3.49313     0.09408    37.13   < 2e-16
happypretty   1.18211     0.07672    15.41   < 2e-16
happyvery     0.52957     0.08460     6.26  3.86e-10
heavenyes     1.74920     0.07739    22.60   < 2e-16
---
Residual deviance:    0.89111  on 2  degrees of freedom
------------------------------------------------------------------------------
```

The deviance is 0.89 ($df = 2$), so the model fits well. For the constraints used, $\lambda_1^Y = 0$ and $\lambda_2^Y = 1.749$. The estimated odds of belief in heaven was $\exp(1.749) = 5.75$ for each happiness level.

7.1.4 Saturated Model for Two-Way Contingency Tables

Categorical variables that have an association rather than being statistically independent satisfy the more complex loglinear model,

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}.$$

The $\{\lambda_{ij}^{XY}\}$ parameters are association terms. The parameters represent interactions between X and Y, whereby the effect of either variable on the expected cell count depends on the category of the other variable.

Direct relationships exist between log odds ratios and the $\{\lambda_{ij}^{XY}\}$. For example, this model for 2×2 contingency tables has the log odds ratio

$$\begin{aligned}\log \theta &= \log\left(\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}\right) = \log \mu_{11} + \log \mu_{22} - \log \mu_{12} - \log \mu_{21} \\ &= (\lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\lambda + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}) \\ &\quad - (\lambda + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) - (\lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}) \\ &= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}.\end{aligned}$$

In $r \times c$ contingency tables, only $(r-1)(c-1)$ association parameters are nonredundant. For example, R specifies the parameters so that the ones in the first row and in the first column are zero, so that $\log \theta$ simplifies to λ_{22}^{XY}. The nonzero parameters are coefficients of cross-products of $(r-1)$ indicator variables for X with $(c-1)$ indicator variables for Y. Tests of independence analyze whether these $(r-1)(c-1)$ parameters equal zero, so the chi-squared tests have residual $df = (r-1)(c-1)$.

This model has a single constant parameter (λ), $(r-1)$ nonredundant $\{\lambda_i^X\}$, $(c-1)$ nonredundant $\{\lambda_j^Y\}$, and $(r-1)(c-1)$ nonredundant $\{\lambda_{ij}^{XY}\}$. The total number of parameters equals $1 + (r-1) + (c-1) + (r-1)(c-1) = rc$. The model has as many parameters as observed cell counts. It is the *saturated* loglinear model, having the maximum possible number of parameters. It describes perfectly any set of cell counts, having deviance of 0.

The next R output fits the saturated model with $X =$ happiness (with *not* as the baseline level without its own indicator variable) and $Y =$ belief in heaven:

```
------------------------------------------------------------------------------
> fit2 <- glm(count ~ happy + heaven + happy:heaven, family=poisson,
+             data=HappyHeaven)
> summary(fit2) # saturated loglinear model
                       Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)             3.46574     0.17678   19.60   < 2e-16
happypretty             1.26165     0.20025    6.30  2.97e-10
happyvery               0.46609     0.22552    2.07    0.0388
heavenyes               1.78129     0.19108    9.32   < 2e-16
happypretty:heavenyes  -0.09358     0.21679   -0.43    0.6660
happyvery:heavenyes     0.07378     0.24329    0.30    0.7617
---
Residual deviance: 1.5321e-14 on 0 degrees of freedom # dev. = 0, perfect fit
------------------------------------------------------------------------------
```

The estimated odds ratios are $\exp(-0.094) = 0.91$ for happiness *pretty* and *not*, $\exp(0.074) = 1.08$ for happiness *very* and *not*, and $\exp(-0.094 - (0.074)) = 0.85$ for happiness *pretty* and *very*. Since the independence model fitted well, none of these estimated odds ratios differ significantly from 1.0.

When a model permits interaction, the estimates of the main effect terms depend on the coding scheme used for the higher-order effects, and the interpretation also depends on that scheme. We restrict our interpretations to the highest-order terms for a variable.

7.1.5 Loglinear Models for Three-Way Contingency Tables

With three-way contingency tables, loglinear models can represent various independence and association patterns. Two-factor association terms describe conditional odds ratios between variables.

For cell expected frequencies $\{\mu_{ijk}\}$, consider the loglinear model

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

Since it contains an *XZ* term, it permits association between X and Z, at each category for Y. It also permits a *YZ* association, at each category for X. It does not contain an *XY* term, so this loglinear model specifies independence between X and Y, at each category for Z, that is, *conditional independence*. This model holds when an association between two variables (X and Y) disappears after we adjust for a third variable (Z). We symbolize the model by (XZ, YZ). The symbol lists the highest-order terms in the model for each variable.

Models that delete additional association terms are too simple to fit most data sets well. For instance, the model that contains only single-factor terms, denoted by (X, Y, Z), is called the *mutual independence model*. It treats each pair of variables as independent, both conditionally and marginally. When variables are chosen wisely for a study, this model is rarely appropriate.

A model that permits all three pairs of variables to have conditional associations is

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

For it, we will see that conditional odds ratios between any two variables are the same at each category of the third variable. This is the property of *homogeneous association* (Section 2.7.5). We symbolize this loglinear model, called the *homogeneous association model*, by (XY, XZ, YZ).

The most general loglinear model for three-way tables adds a three-factor interaction term, λ_{ijk}^{XYZ}, to the homogeneous association model. Denoted by (XYZ), it is the saturated model. It provides a perfect fit.

7.1.6 Two-Factor Parameters Describe Conditional Associations

The two-factor parameters in the homogeneous association model relate directly to conditional odds ratios. We illustrate for three-way contingency tables in which X and Y each have two categories. The *XY* conditional odds ratio $\theta_{XY(k)}$ describes the association

between X and Y in partial table k of Z (recall Section 2.7.4). From an argument similar to that in Section 7.1.4,

$$\log \theta_{XY(k)} = \log\left(\frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}\right) = \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}.$$

For software (such as R) that sets the parameter for the first category of each factor equal to 0, this simplfies to λ_{22}^{XY}, and the odds ratio is $\exp(\lambda_{22}^{XY})$ in each partial table.

This log odds ratio between X and Y in partial table k for Z does not depend on k, so it is the same at every category of Z. The model also has equal *XZ* odds ratios at different categories of Y and it has equal *YZ* odds ratios at different categories of X. Any model not having the three-factor term λ_{ijk}^{XYZ} satisfies homogeneous association.

7.1.7 Example: Student Alcohol, Cigarette, and Marijuana Use

Table 7.1 comes from a survey that asked students in their final year of a high school near Dayton, Ohio, whether they had ever used alcohol, cigarettes, or marijuana. Denote the variables in this $2 \times 2 \times 2$ contingency table by A for alcohol use, C for cigarette use, and M for marijuana use.

Table 7.1 Alcohol (A), cigarette (C), and marijuana (M) use for high school seniors.

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Source: Thanks to Prof. Harry Khamis, Wright State University and United Health Services in Dayton for these data (the `Substance` data file at the text website).

Table 7.2 shows fitted values for some loglinear models. The fit for the homogeneous association model (AC, AM, CM) is very close to the observed data, which are the fitted values for the saturated model (ACM). The simpler models fit poorly.

Table 7.2 Fitted values for loglinear models applied to Table 7.1.

Alcohol Use	Cigarette Use	Marijuana Use	Loglinear Model			
			(A, C, M)	(AM, CM)	(AC, AM, CM)	(ACM)
Yes	Yes	Yes	540.0	909.24	910.4	911
		No	740.2	438.84	538.6	538
	No	Yes	282.1	45.76	44.6	44
		No	386.7	555.16	455.4	456
No	Yes	Yes	90.6	4.76	3.6	3
		No	124.2	142.16	42.4	43
	No	Yes	47.3	0.24	1.4	2
		No	64.9	179.84	279.6	279

Table 7.3 Estimated odds ratios for loglinear models fitted to Table 7.1.

	Conditional Association			Marginal Association		
Model	*AC*	*AM*	*CM*	*AC*	*AM*	*CM*
(A, C, M)	1.0	1.0	1.0	1.0	1.0	1.0
(AM, CM)	1.0	61.9	25.1	2.7	61.9	25.1
(AC, AM, CM)	7.8	19.8	17.3	17.7	61.9	25.1
(ACM) Level 1	13.8	24.3	17.5	17.7	61.9	25.1
(ACM) Level 2	7.7	13.5	9.7			

Table 7.3 illustrates association patterns for these models by presenting estimated marginal and conditional odds ratios. A marginal odds ratio ignores the third factor, whereas the conditional odds ratio adjusts for it. For example, the entry 1.0 for the *AC* conditional odds ratio for model (AM, CM) is the common value of the *AC* fitted odds ratios at each category of M,

$$1.0 = \frac{909.24 \times 0.24}{45.76 \times 4.76} = \frac{438.84 \times 179.84}{555.16 \times 142.16}.$$

This model implies conditional independence between alcohol use and cigarette use, given marijuana use. Conditional odds ratios equal 1.0 for each pairwise term not appearing in a model, such as the *AC* association in model (AM, CM). The entry 2.7 for the *AC marginal* association for this model is the odds ratio for the marginal *AC* fitted table, collapsed over M,

$$2.7 = \frac{(909.24 + 438.84)(0.24 + 179.84)}{(45.76 + 555.16)(4.76 + 142.16)}.$$

This marginal odds ratio differs from 1.0. Recall that conditional independence does not imply marginal independence. The odds ratios for the observed data are those reported for the saturated model (ACM).

Some conditional odds ratios in Table 7.3 equal corresponding marginal odds ratios. Section 7.4.2 presents a condition that guarantees this. This equality does not normally happen for loglinear models that contain all the pairwise associations.

For model (AC, AM, CM) or simpler models, we can estimate a conditional odds ratio using the parameter estimates. For example, the estimated conditional *AC* odds ratio is

$$\exp(\hat{\lambda}_{11}^{AC} + \hat{\lambda}_{22}^{AC} - \hat{\lambda}_{12}^{AC} - \hat{\lambda}_{21}^{AC}).$$

Software constrains parameters so that three of these four terms are 0. For example, from the R output shown next, the estimated conditional *AC* odds ratio is $\exp(2.0545) = 7.8$:

```
------------------------------------------------------------------------------
> Drugs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Substance.dat",
+                     header=TRUE)
> Drugs
  alcohol cigarettes marijuana count # data file has 8 rows, for 8 cell counts
```

```
1     yes        yes       yes   911
...
8      no         no        no   279
> A <- Drugs$alcohol; C <- Drugs$cigarettes; M <- Drugs$marijuana
> fit <- glm(count ~ A + C + M + A:C + A:M + C:M, family=poisson, data=Drugs)
> summary(fit) # homogeneous association model
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  5.6334      0.0597    94.36   < 2e-16
Ayes         0.4877      0.0758     6.44  1.22e-10
Cyes        -1.8867      0.1627   -11.60   < 2e-16
Myes        -5.3090      0.4752   -11.17   < 2e-16
Ayes:Cyes    2.0545      0.1741    11.80   < 2e-16 # AC log odds ratio = 2.0545
Ayes:Myes    2.9860      0.4647     6.43  1.31e-10
Cyes:Myes    2.8479      0.1638    17.38   < 2e-16
---
Residual deviance:    0.37399  on 1  degrees of freedom
--------------------------------------------------------------------------------
```

7.2 STATISTICAL INFERENCE FOR LOGLINEAR MODELS

Table 7.3 shows that estimates of conditional and marginal odds ratios are highly dependent on the model. This highlights the importance of good model selection. An estimate from this table is informative only if its model fits well. This section shows how to check goodness of fit, conduct inference, and extend loglinear models to higher dimensions.

7.2.1 Chi-Squared Goodness-of-Fit Tests

The deviance statistic for logistic models with categorical explanatory variables (Section 5.2.2) can also assess goodness of fit of loglinear models by comparing the cell fitted values to the observed counts. The chi-squared degrees of freedom equal the number of cell counts minus the number of model parameters. The deviance is the likelihood-ratio statistic for testing that all parameters that are in the saturated model but not in the working model equal 0.

For the student substance-use survey (Table 7.1), Table 7.4 shows the deviance for some loglinear models. The model (AC, AM, CM) that permits all pairwise associations but assumes homogeneous association fits well ($P = 0.54$). The models that take out any association term fit poorly, having P-values below 0.001.

Table 7.4 Goodness-of-fit tests for loglinear models relating alcohol (A), cigarette (C), and marijuana (M) use.

Model	Deviance	*df*	P-value*
(AC, AM)	497.4	2	< 0.001
(AC, CM)	92.0	2	< 0.001
(AM, CM)	187.8	2	< 0.001
(AC, AM, CM)	0.4	1	0.54

7.2.2 Cell Standardized Residuals for Loglinear Models

Goodness-of-fit statistics merely provide global indications of lack of fit. Cell residuals show the quality of fit cell by cell. Sometimes they indicate that certain cells display lack of fit in a model that otherwise fits well. Section 2.4.5 introduced standardized residuals for the independence model and Section 3.4.5 presented them generally for GLMs. They divide differences between observed and fitted counts by their standard errors. When the model holds and the expected frequencies are relatively large, standardized residuals have approximately a standard normal distribution.

The next R output shows standardized residuals for loglinear models (AC, AM, CM) and (AM, CM):

```
------------------------------------------------------------------------------
> fit <- glm(count ~ A + C + M + A:C + A:M + C:M, family=poisson, data=Drugs)
> fit2 <- glm(count ~ A + C + M + A:M + C:M, family=poisson, data=Drugs)
> deviance(fit); deviance(fit2)
[1] 0.3739859
[1] 187.7543
> res <- rstandard(fit,type="pearson"); res2 <- rstandard(fit2,type="pearson")
> data.frame(A, C, M, Drugs$count, fitted(fit), res, fitted(fit2), res2)
    A   C   M count  fitted.fit.    res  fitted.fit2.    res2
1 yes yes yes   911     910.383   0.633      909.240    3.696
2 yes yes  no   538     538.617  -0.633      438.840   12.805
3 yes  no yes    44      44.617  -0.633       45.760   -3.696
4 yes  no  no   456     455.383   0.633      555.160  -12.805
5  no yes yes     3       3.617  -0.633        4.760   -3.696
6  no yes  no    43      42.383   0.633      142.160  -12.805
7  no  no yes     2       1.383   0.633        0.240    3.696
8  no  no  no   279     279.617  -0.633      179.840   12.805
------------------------------------------------------------------------------
```

The model (AM, CM) of *AC* conditional independence has $df = 2$ for testing fit. The two nonredundant standardized residuals, which have magnitude 3.70 and 12.80, refer to checking *AC* independence at each category of M. The large residuals reflect the overall poor fit. At each category for M, the large positive standardized residuals occur when A and C are both *yes* or both *no*. More of these students have used both or neither of alcohol and cigarettes than one would expect if their usage were conditionally independent. For model (AC, AM, CM), since $df = 1$, only one standardized residual is nonredundant, having magnitude 0.63. The model's residual deviance is small, so these indicate a good fit.

Two caveats: When a table has a large number of cells, some standardized residuals may be large merely by chance. When n is extremely large, the standardized residuals (like any test statistic) may be large even though, in practical terms, the model fit is adequate.

7.2.3 Significance Tests about Conditional Associations

To test a conditional association in a model, we compare the model to the simpler model not containing that association. The likelihood-ratio statistic for testing that a model term equals zero is identical to the difference between the deviances for the models.

For example, for model (AC, AM, CM), the null hypothesis of conditional independence between alcohol use and cigarette smoking states that the λ^{AC} term equals zero. The test analyzes whether the simpler model (AM, CM) of *AC* conditional independence holds, against the alternative that model (AC, AM, CM) holds. From the R output in Section 7.2.2 (and Table 7.4), this test statistic equals $187.75 - 0.37 = 187.38$. It has $df = 2 - 1 = 1$ ($P < 0.0001$). This is strong evidence of an *AC* conditional association. Likelihood-ratio tests also provide strong evidence of *AM* and *CM* conditional associations:

```
-------------------------------------------------------------------------------
> library(car)
> Anova(fit) # likelihood-ratio tests for pairwise conditional associations
    LR Chisq  Df  Pr(>Chisq)
A:C   187.38   1   < 2.2e-16
A:M    91.64   1   < 2.2e-16
C:M   497.00   1   < 2.2e-16
-------------------------------------------------------------------------------
```

7.2.4 Confidence Intervals for Conditional Odds Ratios

For loglinear models in which the highest-order terms are two-factor associations, the estimates refer to conditional log odds ratios. We can construct confidence intervals for the true values.

We illustrate for the association between alcohol and marijuana use, for model (AC, AM, CM). From the output in Section 7.2.3, partly repeated below, the estimated AM conditional log odds ratio is 2.986, with $SE = 0.465$. A 95% Wald confidence interval for the true conditional odds ratio is $\exp[2.986 \pm 1.96(0.465)]$, which is (8.0, 49.2). At each category of C, students who have smoked marijuana have estimated odds of having drunk alcohol that are between 8.0 and 49.2 times the estimated odds for students who have not smoked marijuana. This and the corresponding profile likelihood confidence interval of (8.8, 56.6) indicate a strong positive conditional association.

```
-------------------------------------------------------------------------------
> fit <- glm(count ~ A + C + M + A:C + A:M + C:M, family=poisson, data=Drugs)
> summary(fit) # showing only log odds ratio estimates
           Estimate  Std. Error  z value  Pr(>|z|)
Ayes:Cyes   2.05453     0.17406   11.803   < 2e-16
Ayes:Myes   2.98601     0.46468    6.426  1.31e-10 # AM log odds ratio = 2.986
Cyes:Myes   2.84789     0.16384   17.382   < 2e-16
---
> exp(confint(fit))
                   2.5%      97.5%
Ayes:Cyes   5.60145e+00   11.09715
Ayes:Myes   8.81405e+00   56.64360 # profile likelihood CI for conditional OR
Cyes:Myes   1.26458e+01   24.06925
-------------------------------------------------------------------------------
```

The 95% profile likelihood confidence intervals are (5.6, 11.1) for the *AC* conditional odds ratio and (12.6, 24.1) for the *CM* conditional odds ratio. These associations also are strong. In summary, this model reveals strong positive conditional associations for each

pair of substances. There is a strong tendency for users of one substance to be users of a second, and this is true both for users and for nonusers of the third. Table 7.3 shows that estimated marginal associations are even stronger. Controlling for the outcome for one substance moderates the association somewhat between the other two.

These analyses pertain to association structure. A different analysis pertains to comparing marginal distributions, for instance to determine if one substance has more usage than the others. Sections 8.1 and 9.2 present that type of analysis.

7.2.5 Bayesian Fitting of Loglinear Models

Alternatively, we can use a Bayesian approach to fitting loglinear models. For the homogeneous association model, we use diffuse prior distributions that treat the loglinear model parameters as independent normal random variables with means of 0 and standard deviations of 10:

```
-------------------------------------------------------------------------------
> library(MCMCpack) # b0 = prior mean, B0 = prior precision = 1/variance
> fitBayes <- MCMCpoisson(count ~ A + C + M + A:C + A:M + C:M, b0=0, B0=0.01,
+                         mcmc=10000000, data=Drugs)
> summary(fitBayes)
1. Empirical mean and standard deviation # showing only association parameters
               Mean       SD
Ayes:Cyes    2.0644  0.17481
Ayes:Myes    3.0639  0.48161
Cyes:Myes    2.8569  0.16442
2. Quantiles for each variable:
              2.5%      25%     50%     75%   97.5%
Ayes:Cyes    1.7296  1.9451  2.0616  2.1806  2.4152
Ayes:Myes    2.2111  2.7274  3.0321  3.3655  4.0983
Cyes:Myes    2.5437  2.7444  2.8540  2.9660  3.1879

> mean(fitBayes[,6] < 0) # posterior prob. that AM log odds ratio < 0
[1] 0                    # (parameter 6 in model is AM log odds ratio)
-------------------------------------------------------------------------------
```

Inferences about the model parameters are substantively the same as with the ML frequentist analysis. For example, the 95% posterior interval for the conditional AM odds ratio is $(\exp(2.211), \exp(4.098))$, which is (9.1, 60.2), compared to the profile likelihood interval of (8.8, 56.6). The posterior probability that the AM log odds ratio is negative is < 0.0001, as is the one-sided P-value for the likelihood-ratio or Wald test of no AM effect against the alternative of a positive one.

7.2.6 Loglinear Models for Higher-Dimensional Contingency Tables

Basic concepts for loglinear models with three-way contingency tables extend readily to multi-way tables. Loglinear models that have two-factor highest-order terms have homogenous associations: two variables that are conditionally associated have the same odds ratios at each combination of levels of the other variables. An absence of a two-factor term implies

conditional independence for those variables. A three-factor term allows heterogeneous conditional associations, whereby the association between any pair of those three variables varies across categories of the third variable, at any settings of the other variables.

For model selection, an exploratory approach first fits the model having only single-factor terms, the model that adds all two-factor terms, and the model that adds all three-factor terms. Fitting such models often reveals a restricted range of good-fitting models.

7.2.7 Example: Automobile Accidents and Seat Belts

Table 7.5 shows results of accidents in the state of Maine for 68,694 passengers in autos and light trucks. The table classifies passengers by gender (G), location of accident (L), seat-belt use (S), and injury (I). The table reports the sample proportion of passengers who were injured. For each *GL* combination, the proportion of injuries was about halved for passengers wearing seat-belts.

Table 7.5 Injury (I) by gender (G), location (L), and seat-belt use (S), with fit of loglinear model (GLS, GI, LI, SI).

		Seat	Injury		(*GLS, GI, LI, SI*)		Sample
Gender	Location	Belt	No	Yes	No	Yes	Prop. Yes
Female	Rural	No	3246	973	3254.7	964.3	0.23
		Yes	6134	757	6093.5	797.5	0.11
	Urban	No	7287	996	7273.2	1009.8	0.12
		Yes	11587	759	11632.6	713.4	0.06
Male	Rural	No	6123	1084	6150.2	1056.8	0.15
		Yes	6693	513	6697.6	508.4	0.07
	Urban	No	10381	812	10358.9	834.1	0.07
		Yes	10969	380	10959.2	389.8	0.03

Source: I am grateful to Dr. Cristanna Cook, Medical Care Development, Augusta, Maine, for supplying these data, which are in the `Accidents2` data file at the text website.

Table 7.6 displays tests of fit for several loglinear models. To investigate the complexity of model needed, we first fitted the model with only single-factor terms, the model containing also all the two-factor terms, and the model containing also all the three-factor terms. Model (G, L, S, I), which implies mutual independence of the four variables, fits very poorly. Model (GL, GS, LS, GI, LI, SI) fits much better but still has lack of fit.

Table 7.6 Goodness-of-fit tests for loglinear models relating injury (I), gender (G), location (L), and seat-belt use (S).

Model	Deviance	*df*	P-value	AIC
(G, L, S, I)	2792.8	11	< 0.0001	2956.2
(GL, GS, LS, GI, LI, SI)	23.4	5	< 0.001	198.8
(GLS, GLI, GSI, LSI)	1.3	1	0.25	184.8
(GLS, GI, LI, SI)	7.5	4	0.11	184.9
(GLI, GS, LS, SI)	18.6	4	0.001	196.0
(GSI, GL, LS, LI)	22.8	4	< 0.001	200.3
(LSI, GL, GS, GI)	20.6	4	< 0.001	198.1

Model (GLS, GLI, GSI, LSI) fits well (deviance $= 1.3$, $df = 1$). This suggests investigating models that are more complex than (GL, GS, LS, GI, LI, SI) but simpler than (GLS, GLI, GSI, LSI).

Table 7.6 shows results of adding a single three-factor term to model (GL, GS, LS, GI, LI, SI). Of the four possible models, (GLS, GI, LI, SI) fits best. According to AIC, it has a similar fit as the model with all the three-factor terms. Table 7.5 displays its fit.

7.2.8 Interpreting Three-Factor Interaction Terms

For model (GLS, GI, LI, SI), each pair of variables is conditionally dependent, and at each category of I the association between G and L or between G and S or between L and S varies across the categories of the remaining variable. The following R output shows its parameter estimates:

```
------------------------------------------------------------------------------
> Accidents <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                          Accidents2.dat", header=TRUE)
> Accidents # 16 cell counts in the contingency table
   gender location seatbelt injury count
1  female    rural       no     no  3246
...
16   male    urban      yes    yes   380
> G <- Accidents$gender; L <- Accidents$location
> S <- Accidents$seatbelt; I <- Accidents$injury
> fit <- glm(count ~ G*L*S + G*I + L*I + S*I, family=poisson, data=Accidents)
> summary(fit) # e.g. G*I represents G + I + G:I
Coefficients: # not showing intercept and main effect estimates
                  Estimate  Std. Error   z value   Pr(>|z|)
Gmale:Lurban      -0.28274     0.02441   -11.584    < 2e-16
Gmale:Syes        -0.54186     0.02590   -20.925    < 2e-16
Lurban:Syes       -0.15752     0.02441    -6.453   1.09e-10
Gmale:Iyes        -0.54483     0.02727   -19.982    < 2e-16 # GI log odds ratio
Lurban:Iyes       -0.75806     0.02697   -28.105    < 2e-16 # LI log odds ratio
Syes:Iyes         -0.81710     0.02765   -29.551    < 2e-16 # SI log odds ratio
Gmale:Lurban:Syes  0.12858     0.03228     3.984   6.78e-05
---
Residual deviance:     7.4645  on  4  degrees of freedom
------------------------------------------------------------------------------
```

Table 7.7 reports the model-based estimated odds ratios. For comparison, the table also shows the results for the model that assumes homogeneous conditional odds ratios for each pair of variables. We can obtain the log odds ratios directly from loglinear parameter estimates. Because I does not occur in a three-factor term, the conditional odds ratio between I and each variable is the same at each combination of categories of the other two variables. From the output, the estimated *SI* conditional odds ratio is $\exp(-0.8171) = 0.44$, with 95% Wald confidence interval $\exp[-0.8171 \pm 1.96(0.02765)]$, or (0.42, 0.47). The odds of injury for passengers wearing seat-belts were less than half the odds for passengers not wearing them, for each gender-location combination. The sample size is very large and

Table 7.7 Estimated conditional odds ratios for two loglinear models.

	Loglinear Model	
Odds Ratio	(GL, GS, LS, GI, LI, SI)	(GLS, GI, LI, SI)
GI	0.58	0.58
LI	0.47	0.47
SI	0.44	0.44
GL (S=no)	0.81	0.75
GL (S=yes)	0.81	0.86
GS (L=urban)	0.63	0.66
GS (L=rural)	0.63	0.58
LS (G=female)	0.92	0.85
LS (G=male)	0.92	0.97

the estimates of odds ratios are precise. The *LI* and *GI* fitted odds ratios in Table 7.7 suggest that, other factors being fixed, injury was less likely in urban than rural accidents and less likely for males than females.

It is inappropriate to interpret the *GL*, *GS*, and *LS* two-factor terms on their own. For example, the presence of the *GLS* term implies that the *GS* odds ratio varies across the categories of L. When a model has a three-factor term, to study the interaction, calculate fitted odds ratios between two variables at each category of the third. Do this at any categories of remaining variables not involved in the interaction. The bottom six lines of Table 7.7 illustrates for model (GLS, GI, LI, SI). For example, the fitted *GS* odds ratio of 0.66 for ($L =$ urban) refers to four fitted values for urban accidents, both the four with (injury $=$ no) and the four with (injury $=$ yes). You can find these directly from the parameter estimates shown in the output, as $\exp(-0.542 + 0.129)$. The fitted *GS* odds ratio for ($L =$ rural) is $\exp(-0.542) = 0.58$. Similarly, the fitted *GL* odds ratio is $\exp(-0.283) = 0.75$ for ($S =$ no) and $\exp(-0.283 + 0.129) = 0.86$ for ($S =$ yes).

7.2.9 Statistical Versus Practical Significance: Dissimilarity Index

The sample size can strongly influence results of any inferential procedure. We are more likely to detect an effect as n increases. This suggests a cautionary remark. For small n, reality may be more complex than indicated by the simplest model that passes a goodness-of-fit test. By contrast, for large n, statistically significant effects can be weak and unimportant.

Table 7.6 indicated that model (GLS, GI, LI, SI) fits much better than (GL, GS, LS, GI, LI, SI): the difference in deviance values is $23.4 - 7.5 = 15.9$, based on $df = 5 - 4 = 1$ ($P = 0.0001$) and its AIC value is considerably smaller. The fitted odds ratios in Table 7.7, however, show that the three-factor interaction is weak. The fitted odds ratio between any two of G, L, and S is similar at both levels of the third variable. The significantly better fit of model (GLS, GI, LI, SI) mainly reflects the enormous sample size. Although the three-factor interaction is weak, it is significant because the large sample provides small standard errors. A comparison of fitted odds ratios for the two models suggests that the simpler model (GL, GS, LS, GI, LI, SI) is adequate for most practical purposes. With a very large sample size, analyses that are affected by n, such as goodness-of-fit tests and AIC, should not be the only criteria for selecting a model.

With large n, it is helpful to summarize the closeness of a model fit to the sample data in a way that, unlike a test statistic, is not affected by n. For a contingency table of arbitrary

dimensions with cell counts $\{n_i = np_i\}$ and fitted values $\{\hat{\mu}_i = n\hat{\pi}_i\}$, one such measure is the *dissimilarity index*,

$$D = \sum |n_i - \hat{\mu}_i|/2n = \sum |p_i - \hat{\pi}_i|/2.$$

This index represents the proportion of sample cases that must move to different cells for the model to achieve a perfect fit. It takes values between 0 and 1, with smaller values representing a better fit. It helps indicate whether the lack of fit is important in a practical sense. A very small D value suggests that the sample data follow the model pattern closely, even though the model is not perfect.

For Table 7.5, model (GL, GS, LS, GI, LI, SI) has $D = 0.0082$ and model (GLS, GI, LI, SI) has $D = 0.0025$. For either model, moving less than 1% of the data yields a perfect fit. The relatively large deviance for model (GL, GS, LS, GI, LI, SI) indicated that the model does not truly hold. Nevertheless, the small value for D suggests that, in practical terms, the model provides a decent fit. The next R output shows the calculation of D:

```
-------------------------------------------------------------------------------
> fit <- glm(count ~ G*L*S + G*I + L*I + S*I, family=poisson, data=Accidents)
> sum(abs(Accidents$count - fitted(fit)))/(2*sum(Accidents$count))
[1] 0.00251  # dissimilarity index for loglinear model (GLS, GI, LI, SI)

> fit2 <- glm(count ~ G*L+G*S+L*S+G*I+L*I+S*I,family=poisson, data=Accidents)
> sum(abs(Accidents$count - fitted(fit2)))/(2*sum(Accidents$count))
[1] 0.00822  # dissimilarity index for model (GL, GS, LS, GI, LI, SI)
-------------------------------------------------------------------------------
```

7.3 THE LOGLINEAR – LOGISTIC MODEL CONNECTION

Loglinear models for contingency tables focus on associations between categorical response variables. Logistic regression models, on the other hand, describe the effects of a set of explanatory variables on a categorical response variable. Though the model types seem distinct, connections exist between them. For a loglinear model, one can construct logits for one response to help interpret the model. Moreover, logistic models for which all the explanatory variables are categorical have equivalent loglinear models.

7.3.1 Using Logistic Models to Interpret Loglinear Models

To understand implications of a loglinear model formula, it can help to form a logit for one of the variables. We illustrate with the homogeneous association model for three-way tables,

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

Suppose Y is binary. When we treat X and Z as explanatory, from an argument similar to that in Section 7.1.4, the model implies that

$$\text{logit}[P(Y = 1)] = \alpha + \beta_i^X + \beta_k^Z,$$

with $\alpha = (\lambda_1^Y - \lambda_2^Y)$, $\beta_i^X = (\lambda_{i1}^{XY} - \lambda_{i2}^{XY})$, and $\beta_k^Z = (\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ})$.

Section 4.3.3 discussed this model, in which the logit depends on X and Z in an additive manner. When Y is binary, the loglinear model of homogeneous association is equivalent to this logistic regression model with factor main effects and no interaction term between them.

7.3.2 Example: Auto Accident Data Revisited

For the data on Maine auto accidents (Table 7.5), we observed that loglinear model (GLS, GI, LI, SI) fits well. That model has the formula

$$\log \mu_{gi\ell s} = \lambda + \lambda_g^G + \lambda_\ell^L + \lambda_s^S + \lambda_i^I + \lambda_{g\ell}^{GL} + \lambda_{gs}^{GS} + \lambda_{\ell s}^{LS} + \lambda_{gi}^{GI} + \lambda_{\ell i}^{LI} + \lambda_{si}^{SI} + \lambda_{g\ell s}^{GLS}.$$

If we treat injury (I) as a response variable and gender (G), location (L), and seat-belt use (S) as explanatory variables, this model implies a logistic model of the form

$$\text{logit}[P(I=1)] = \alpha + \beta_g^G + \beta_\ell^L + \beta_s^S. \tag{7.2}$$

Here, G, L, and S are all associated with I, but without interacting.

Odds ratios relate to two-factor loglinear parameters and main-effect logistic parameters. For instance, in model (7.2), the log odds ratio for the effect of S on I equals $\beta_1^S - \beta_2^S$. This equals $\lambda_{11}^{SI} + \lambda_{22}^{SI} - \lambda_{12}^{SI} - \lambda_{21}^{SI}$ in the loglinear model. These values are the same no matter how software sets up constraints for the parameters. For example, $\hat{\beta}_1^S - \hat{\beta}_2^S = -0.817$ for model (7.2) and $\hat{\lambda}_{11}^{SI} + \hat{\lambda}_{22}^{SI} - \hat{\lambda}_{12}^{SI} - \hat{\lambda}_{21}^{SI} = -0.817$ for model (GLS, GI, LI, SI).

Loglinear models are GLMs that treat the 16 cell counts in Table 7.5 as outcomes of 16 Poisson variates. Logistic models are GLMs that treat the table as outcomes of 8 binomial variates giving injury counts at the 8 possible settings for the variables G, L, and S. Although the sampling models differ, the results from fits of corresponding models are identical. The fitted values, deviance, residual *df*, and standardized residuals for logistic model (7.2) are identical to those in Tables 7.5 to 7.7 for the loglinear model (GLS, GI, LI, SI).

The output in Section 7.2.8 showed R code for fitting the loglinear model to a data file of 16 observations for which the variable *count* is treated as having a Poisson distribution. Following is output for the logistic model fitted to a data file of 8 observations in which the number having an injury is treated as having a binomial distribution:

```
------------------------------------------------------------------------------
> Injury <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Injury_binom.dat",
+                      header=TRUE) # Injury_binom data file at text website
> Injury
  gender location seatbelt    no  yes
1 female    urban       no  7287  996
2 female    urban      yes 11587  759
... # 8 lines in data file, one for each binomial on injury, given (G, L, S)
8   male    rural      yes  6693  513
> G <- Injury$gender; L <- Injury$location; S <- Injury$seatbelt
> fit2 <- glm(yes/(no+yes) ~ G + L + S, family=binomial, weights=no+yes,
+             data=Injury)
> summary(fit2)
             Estimate  Std. Error  z value  Pr(>|z|)
```

```
(Intercept)  -1.21640     0.02649   -45.92    <2e-16
Gmale        -0.54483     0.02727   -19.98    <2e-16
Lurban       -0.75806     0.02697   -28.11    <2e-16
Syes         -0.81710     0.02765   -29.55    <2e-16
---
Residual deviance:    7.4645  on 4  degrees of freedom
-------------------------------------------------------------------------
```

The log odds ratios relating to an association with I in the loglinear model output are the same as those with I as a response variable in the logistic model ouput. Also, the deviances are identical, 7.46 ($df = 4$) in each case.

7.3.3 Condition for Equivalent Loglinear and Logistic Models

For a particular logistic model containing only categorical factors as explanatory variables for a response variable Y, the loglinear model that has the same fit is the one that has the same association structure between those explanatory variables and Y and that contains a general interaction term for relationships among the explanatory variables. The logistic model does not describe relationships among explanatory variables, so it assumes nothing about their association structure.

The logistic model $\text{logit}[P(I = 1)] = \alpha + \beta_g^G + \beta_\ell^L + \beta_s^S$ for a four-way table contains main effect terms for the explanatory variables, but no interaction terms. This model corresponds to the loglinear model that contains the fullest interaction term among the explanatory variables and associations between each explanatory variable and the response I, namely model (GLS, GI, LI, SI). When a response variable has more than two categories, relevant loglinear models correspond to baseline-category logit models (Section 6.1).

7.3.4 Loglinear/Logistic Model Selection Issues

When a study has a single categorical response variable, it is more sensible to fit logistic models directly, rather than loglinear models. Indeed, equation (7.2) shows how much simpler the logistic structure is. The loglinear approach is better suited for cases with more than one response variable, as in studying association patterns for the subtance use example in Section 7.1.7.

When certain marginal totals are fixed by the sampling design or by the response–explanatory distinction, the model should contain the term for that margin. This is because the ML fit forces the corresponding fitted totals to be identical to those marginal totals. To illustrate, suppose we treat the counts $\{n_{g+\ell+}\}$ in Table 7.5 as fixed at each combination of levels of $G =$ gender and $L =$ location. Then a loglinear model should contain the *GL* two-factor term, because this ensures that $\{\hat{\mu}_{g+\ell+} = n_{g+\ell+}\}$. That is, the model should be at least as complex as model (GL, S, I). If 20,629 women had accidents in urban locations, then the fitted counts have 20,629 women in urban locations. Related to this point, the modeling process should concentrate on terms linking response variables and terms linking explanatory variables to response variables. Allowing a general interaction term among the explanatory variables has the effect of fixing totals at combinations of their levels. If G and

L are both explanatory variables, models assuming conditional independence between G and L are not of interest.

For Table 7.5, I is a response variable and S might be treated either as a response or explanatory variable. If it is explanatory, we treat the $\{n_{g+\ell s}\}$ totals as fixed and fit logistic models for the I response. If S is also a response, we consider the $\{n_{g+\ell+}\}$ totals as fixed and consider loglinear models that are at least as complex as (GL, S, I). Such models focus on the effects of G and L on S and on I as well as the association between S and I.

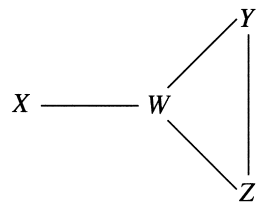
7.4 INDEPENDENCE GRAPHS AND COLLAPSIBILITY

We next present a graphical representation for conditional independences in loglinear models. The graph indicates which variables need to be controlled for a conditional independence to occur. This representation is also helpful for revealing implications of models, such as determining when marginal and conditional odds ratios are identical.

7.4.1 Independence Graphs

An *independence graph* for a loglinear model has a set of vertices, each vertex representing a variable. There are as many vertices as dimensions of the contingency table. Any two vertices can be connected by an edge. A missing edge between two vertices represents conditional independence between the corresponding two variables. The graph is *undirected*, with the edges not pointing from one variable to another.

For example, for a four-way contingency table, the loglinear model (WX, WY, WZ, YZ) lacks *XY* and *XZ* association terms. It assumes that X and Y are independent and that X and Z are independent, conditional on the other two variables. The independence graph



portrays this model. Edges do not connect X with Y or X with Z, because those two pairs are conditionally independent.

A *path* in an independence graph is a sequence of edges leading from one variable to another. Two variables X and Y are said to be *separated* by a subset of variables if all paths connecting X and Y intersect that subset. In the above independence graph, W separates X and Y, since any path connecting X with Y goes through W. The subset $\{W, Z\}$ also separates X and Y. A fundamental result on conditional independences for undirected graphs states:

Conditional independence and separation*: Two variables are conditionally independent given any subset of variables that separates them.*

Thus, not only are X and Y conditionally independent given W and Z, but also given W alone. Similarly, X and Z are conditionally independent given W alone.

The loglinear model (WX, XY, YZ) has the independence graph

$$W \text{\_\_\_\_\_} X \text{\_\_\_\_\_} Y \text{\_\_\_\_\_} Z.$$

Here, W and Z are conditionally independent given X alone or given Y alone or given both X and Y. Also, W and Y are conditionally independent, given X alone or given X and Z, and X and Z are conditionally independent, given Y alone or given Y and W.

7.4.2 Collapsibility Conditions for Contingency Tables

Sometimes researchers collapse multiway contingency tables to make them simpler to describe and analyze. However, marginal associations in collapsed tables may differ from conditional associations. For example, if X and Y are conditionally independent, given Z, they are not usually marginally independent. Under the following *collapsibility conditions*, odds ratios for a model *are* identical in partial tables as in the marginal table:

> ***Collapsibility of three-way contingency tables****: XY marginal and conditional odds ratios are identical if either Z and X are conditionally independent or if Z and Y are conditionally independent.*

The conditions state that the variable treated as the control (Z) is conditionally independent of X or Y, or both. These conditions correspond to loglinear models (XY, YZ) and (XY, XZ). To illustrate, for Table 7.1 with $A =$ alcohol use, $C =$ cigarette use, and $M =$ marijuana use, the model (AM, CM) of AC conditional independence has the independence graph

$$A \text{\_\_\_\_\_} M \text{\_\_\_\_\_} C.$$

For this model, the *AM* conditional odds ratios are the same as the *AM* marginal odds ratio collapsed over C. In fact, from Table 7.3, both the fitted marginal and conditional *AM* odds ratios equal 61.9. Similarly, the *CM* association is collapsible. The *AC* association is not, however. Thus, A and C may be marginally dependent, even though they are conditionally independent in this model. In fact, from Table 7.3, the model's fitted *AC* marginal odds ratio equals 2.7, not 1.0. For the model (AC, AM, CM) of homogeneous association, no pair is conditionally independent, so no collapsibility conditions are fulfilled. For this model, each pair of variables can have quite different fitted marginal and conditional associations.

The collapsibility conditions extend to multiway contingency tables:

> ***Collapsibility for multiway contingency tables****: Suppose that variables partition into three mutually exclusive subsets, A, B, C, such that B separates A and C,*
>
> $$A \text{\_\_\_\_\_} B \text{\_\_\_\_\_} C,$$
>
> *When we collapse the table over the variables in C, model parameters relating variables in A and model parameters relating variables in A with variables in B are unchanged.*

Under this condition, it follows that the corresponding associations are unchanged, as described by odds ratios based on those parameters.

7.4.3 Example: Loglinear Model Building for Student Substance Use

Table 7.1 introduced data on usage of alcohol (A), cigarettes (C), and marijuana (M) by high school students. When we classify the students also by gender (G) and race (R), the five-dimensional contingency table shown in Table 7.8 results. In selecting a model, we treat A, C, and M as response variables and G and R as explanatory variables. Since G and R are explanatory, it does not make sense to estimate association or assume conditional independence for that pair. By remarks in Section 7.3.4, a model should contain the *GR* term, so that the *GR* fitted marginal totals are the same as the corresponding sample marginal totals.

Table 7.8 Alcohol, cigarette, and marijuana use for high school seniors, by gender and race.

		Marijuana Use							
		Race = White				Race = Other			
		Female		Male		Female		Male	
Alcohol Use	Cigarette Use	Yes	No	Yes	No	Yes	No	Yes	No
Yes	Yes	405	268	453	228	23	23	30	19
	No	13	218	28	201	2	19	1	18
No	Yes	1	17	1	17	0	1	1	8
	No	1	117	1	133	0	12	0	17

Source: Professor Harry Khamis, Wright State University; data in `Substance2` data file at text website.

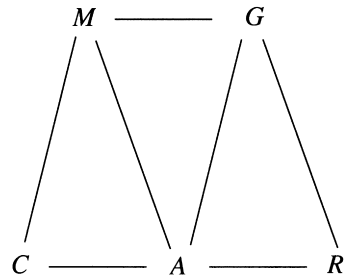
Table 7.9 shows the deviance for some loglinear models. Because many cell counts are small, the chi-squared approximation for the deviance may be poor, but this index is useful for comparing models.

Table 7.9 Goodness-of-fit tests for models relating alcohol (A), cigarette (C), and marijuana (M) use, by gender (G) and race (R).

Model	Deviance	*df*
1. Mutual independence + GR	1325.14	25
2. Homogeneous association	15.34	16
3. All three-factor terms	5.27	6
4. ($AC, AM, CG, CM, AG, AR, GM, GR, MR$)	15.78	17
5. ($AC, AM, CM, AG, AR, GM, GR, MR$)	16.73	18
6. ($AC, AM, CM, AG, AR, GM, GR$)	19.91	19

The first model listed in Table 7.9 contains only the *GR* association and assumes conditional independence for the other nine pairs of associations. It fits very poorly. The homogeneous association model, on the other hand, seems to fit well. The model containing all the three-factor terms also fits well, but the improvement in fit is not great (difference in deviance of 10.07 based on $df = 10$). Thus, we consider models without three-factor terms.

Beginning with the homogeneous association model, we eliminate two-factor terms that do not make significant contributions. We use a backward elimination process, sequentially taking out terms for which the resulting increase in deviance is smallest, when refitting the model. However, we do not delete the *GR* term relating the explanatory variables. Nine pairwise associations are candidates for removal from model (2) and the smallest increase in deviance occurs in removing the *CR* term. After removing that term (model 4), the smallest additional increase results from removing the *CG* term (model 5). After next removing the *MR* term (model 6), additional removals have a more severe effect. It seems safest not to drop additional terms. Model (6) has an independence graph



Consider the sets $\{C\}, \{A, M\}$, and $\{G, R\}$. For this model, every path between C and $\{G, R\}$ involves a variable in $\{A, M\}$. Given the outcome on alcohol use and marijuana use, the model states that cigarette use is independent of both gender and race. Collapsing over the explanatory variables race and gender, the CA and CM conditional associations are the same as with the model (AC, AM, CM) fitted in Section 7.1.7.

7.4.4 Collapsibility and Logistic Models

The collapsibility conditions apply also to logistic models. For example, consider a clinical trial to study the association between a binary response Y and a binary treatment variable, using data from several centers. The model

$$\text{logit}[P(Y = 1)] = \alpha + \beta x + \beta_k$$

with indicator x for treatment and effect β_k for center k assumes that the treatment effect β is the same for each center. Since this model corresponds to loglinear model (XY, XZ, YZ), the estimated treatment effect may differ if we collapse the table over the center factor. The estimated treatment conditional odds ratio, $\exp(\hat{\beta})$, differs from the sample odds ratio in the marginal 2×2 table.

The collapsibility conditions also have implications for quantitative explanatory variables in logistic regression. Suppose x_1 and x_2 are uncorrelated, as in many designed experiments. Then, as mentioned in Section 4.4.5 and unlike in the ordinary linear model, when x_2 is added to a model containing x_1, the effect of x_1 typically changes. It would not change if x_2 were conditionally independent of x_1 rather than merely uncorrelated with it.

For more details about independence graphs and related loglinear models, see Whittaker (1990). Directed graphs and related causal diagrams are beyond the scope of this book.[2]

[2] For an introduction, see S. Greenland, J. Pearl, and J. Robins, *Epidemiology* **10**: 37–48 (1999), available at `www.ncbi.nlm.nih.gov/pubmed/9888278`.

7.5 MODELING ORDINAL ASSOCIATIONS IN CONTINGENCY TABLES

The loglinear models presented so far have a serious limitation: they treat all classifications as nominal. If we change the order of a variable's categories in any way, we get the same fit. For ordinal variables, these models ignore important information.

Table 7.10, from a General Social Survey, illustrates the inadequacy of ordinary loglinear models for analyzing ordinal data. Subjects were asked their opinion about a man and woman having sex relations before marriage. They were also asked whether methods of birth control should be made available to teenagers. Both classifications have ordered categories. The loglinear model of independence has deviance of 127.65 based on $df = 9$. The model fits poorly, yet adding the ordinary association term makes the model saturated and unhelpful.

Table 7.10 Opinions about premarital sex and teenage birth control, showing fitted values for independence model (first parenthesized set) and for linear-by-linear association model (second set).

	Teenage Birth Control			
Premarital Sex	Strongly Disagree	Disagree	Agree	Strongly Agree
Always wrong	81	68	60	38
	(42.4)	(51.2)	(86.4)	(67.0)
	(80.9)	(67.6)	(69.4)	(29.1)
Almost always wrong	24	26	29	14
	(16.0)	(19.3)	(32.5)	(25.2)
	(20.8)	(23.1)	(31.5)	(17.6)
Wrong only sometimes	18	41	74	42
	(30.1)	(36.3)	(61.2)	(47.4)
	(24.4)	(36.1)	(65.7)	(48.8)
Not wrong at all	36	57	161	157
	(70.6)	(85.2)	(143.8)	(111.4)
	(33.0)	(65.1)	(157.4)	(155.5)

Source: General Social Survey; data in `Teenagers` data file at text website.

Table 7.10 also contains fitted values. Observed counts are much larger than the independence model predicts in the corners where both responses are the most negative possible (*always wrong* with *strongly disagree*) or the most positive possible (*not wrong at all* with *strongly agree*). By contrast, observed counts are much smaller than fitted counts in the other two corners. Cross-classifications of ordinal variables often exhibit their greatest deviations from independence in the corner cells. This pattern suggests a positive trend. Subjects who felt more favorable to making birth control available to teenagers also tended to feel more tolerant about premarital sex.

Models for ordinal variables use association terms that permit negative or positive trends. The models are more complex than the independence model yet simpler than the saturated model.

7.5.1 Linear-by-Linear Association Model

To reflect category orderings, an ordinal loglinear model assigns scores $u_1 \leq u_2 \leq \cdots \leq u_r$ to the r rows and $v_1 \leq v_2 \leq \cdots \leq v_c$ to the c columns. In practice, the most common choice is $\{u_i = i\}$ and $\{v_j = j\}$, the row and column numbers. The ordinal model is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j.$$

The $\beta u_i v_j$ term represents the deviation of $\log \mu_{ij}$ from independence. The deviation is linear in the Y scores at a fixed level of X and linear in the X scores at a fixed level of Y. In column j, for instance, the deviation is a linear function of X, having form (slope) $\times$ (score for X), with slope βv_j. Because of this property, this model is called the *linear-by-linear association model* (abbreviated, $L \times L$). This linear-by-linear deviation implies that the model has its greatest departures from independence in the corners of the table. The parameter β specifies the direction and strength of association. When $\beta > 0$, there is a tendency for Y to increase as X increases. The independence model is the special case $\beta = 0$. When $\beta < 0$, there is a tendency for Y to decrease as X increases.

For the 2×2 table created with the four cells intersecting rows a and c with columns b and d, the $L \times L$ model has odds ratio

$$\frac{\mu_{ab}\mu_{cd}}{\mu_{ad}\mu_{cb}} = \exp[\beta(u_c - u_a)(v_d - v_b)]. \tag{7.3}$$

The association is stronger as $|\beta|$ increases and for pairs of categories that are farther apart. The odds ratios formed using adjacent rows and adjacent columns are *local odds ratios*. Figure 7.1 portrays some local odds ratios. For unit-spaced scores such as $\{u_i = i\}$ and $\{v_j = j\}$, the local odds ratios have the common value

$$\frac{\mu_{ab}\mu_{a+1,b+1}}{\mu_{a,b+1}\mu_{a+1,b}} = e^{\beta}.$$

Any set of equally spaced row and column scores has the property of uniform local odds ratios. This special case is called *uniform association*.

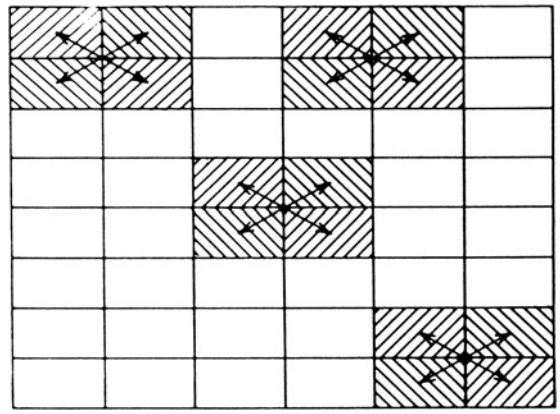


Figure 7.1 Constant local odds ratio implied by the uniform association model.

7.5.2 Example: Linear-by-Linear Association for Sex Opinions

Table 7.10 also reports fitted values for the linear-by-linear association model, using row scores $(1, 2, 3, 4)$ and column scores $(1, 2, 3, 4)$. The deviance for this uniform association version of the model is 11.53, with $df = 8$. Compared to the independence model, the $L \times L$ model provides a dramatic improvement in fit, especially in the corners of the table. Here is R output:

```
------------------------------------------------------------------------------
> Teenagers <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                         Teenagers.dat", header=TRUE)
> Teenagers
   sex birth count
1    1     1    81 # 16 rows, for cell counts in the contingency table
...
16   4     4   157
> fit <- glm(count ~ factor(sex) + factor(birth) + sex:birth, family=poisson,
+            data=Teenagers) # quantitative sex-by-birth interaction term
> summary(fit) # not showing intercept, 3 sex and 3 birth main effects
                Estimate  Std. Error  z value  Pr(>|z|)
sex:birth        0.28584     0.02824   10.122   < 2e-16 # beta estimate
---
Residual deviance:  11.534  on  8  degrees of freedom

> library(car)
> Anova(fit) # likelihood-ratio test for LxL association term
          LR Chisq  Df  Pr(>Chisq)
sex:birth  116.119   1   < 2.2e-16
------------------------------------------------------------------------------
```

The positive ML estimate $\hat{\beta} = 0.286$ suggests that subjects having more favorable attitudes about availability of teen birth control also tend to have more tolerant attitudes about premarital sex. The estimated local odds ratio is $\exp(\hat{\beta}) = \exp(0.286) = 1.33$. The strength of association seems weak. From (7.3), however, fitted odds ratios are stronger for pairs of categories having greater distances between scores. For example, the estimated odds ratio for the four corner cells equals

$$\exp[\hat{\beta}(u_4 - u_1)(v_4 - v_1)] = \exp[0.286(4 - 1)(4 - 1)] = 13.1.$$

To treat categories 2 and 3 as farther apart than categories 1 and 2 or categories 3 and 4, we could instead use scores such as $\{1, 2, 4, 5\}$ for the rows and columns. The $L \times L$ model then has deviance 8.8.

7.5.3 Ordinal Significance Tests of Independence

For the linear-by-linear association model, the hypothesis of independence is H_0: $\beta = 0$. The likelihood-ratio test statistic equals the reduction in deviance between the independence and $L \times L$ models. This statistic refers to a single parameter (β) and has $df = 1$. For Table 7.10, the reduction in deviance of $127.65 - 11.53 = 116.12$ has $P < 0.0001$, extremely strong evidence of an association. The Wald statistic $z^2 = (\hat{\beta}/SE)^2 =$

$(0.2858/0.0282)^2 = 102.4$ also shows strong evidence of a positive trend. The correlation statistic (2.6) for testing independence (Section 2.4.1) is usually similar to the likelihood-ratio and Wald statistics.[3] For Table 7.10, it equals 112.6. All three statistics have $df = 1$.

The linear-by-linear association model generalizes to multiway tables with ordinal variables, such as by a model for three-way tables that has homogeneous linear-by-linear association between X and Y at each category of Z. Chapter 6 presented other ways of using ordinality, based on models that create ordinal logits. To distinguish between an ordinal response variable and explanatory variables, it is more sensible to apply an ordinal logit model than a loglinear model. For further details about loglinear models and their generalizations for nominal and ordinal variables as well as relevant R functions, see Kateri (2014).

7.6 LOGLINEAR MODELING OF COUNT RESPONSE VARIABLES *

Many discrete response variables have *counts* as possible y outcomes. Section 3.3 introduced loglinear models that assume a Poisson distribution for Y. Like counts, Poisson random variables can take any nonnegative integer value. This section shows how to adapt such models to analyze *rates* at which the outcome occurs.

Section 3.3.4 explained that count data often vary more than would be expected if the response distribution truly were Poisson — the phenomenon known as *overdispersion*. The Poisson distribution has variance equal to the mean, but if the variance equals the mean when *all* relevant variables are controlled, it exceeds the mean when only a *subset* of those variables is controlled. This section introduces more flexible loglinear models for count data, using the *negative binomial distribution*. It arises as a type of mixture of Poisson distributions.[4]

7.6.1 Count Regression Modeling of Rate Data

When events occur over time, space, or some other index of size, models can focus on the *rate* at which the events occur. For example, in analyzing numbers of murders in 2019 for a sample of cities, we could form a rate for each city by dividing the number of murders by the city's population size. A model might describe how the rate depends on explanatory variables such as the city's unemployment rate, median income, and percentage of residents having completed high school.

When a response count y has an index (such as population size) equal to t, the sample rate is y/t. The expected value of the rate is μ/t, where $\mu = E(Y)$. A loglinear model for the expected rate has the form

$$\log(\mu/t) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p.$$

This model has equivalent representation

$$\log \mu - \log t = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p.$$

The adjustment term, $-\log t$, to the log of the mean is called an *offset*. Standard GLM software can fit a model having an offset term.

[3] In fact, the correlation statistic is the score statistic for H_0: $\beta = 0$.

[4] For a given mean, if Y has a Poisson distribution, but the means vary according to a *gamma* distribution, unconditionally the distribution is negative binomial.

For this loglinear model, the expected number of outcomes satisfies

$$\mu = t \exp(\alpha + \beta_1 x_1 + \cdots + \beta_p x_p).$$

The mean μ is proportional to the index t, with proportionality constant depending on the values of the explanatory variables. Doubling the population size t, for example, also doubles the expected number of murders μ.

7.6.2 Example: Death Rates for Lung Cancer Patients

Table 7.11 shows survival and death for 539 males diagnosed with lung cancer. The prognostic factors are histology and stage of disease, with observations grouped into two-month intervals of follow-up after the diagnosis. For each cell specifying a particular length of follow-up, histology, and stage of disease, the table shows the number of deaths and the number of months of observations of subjects still alive during that follow-up interval. We treat the death counts in the table as independent Poisson variates.

Table 7.11 Number of deaths from lung cancer, by histology, stage of disease, and follow-up time interval.[a]

Follow-up Time Interval (months)	Disease Stage:	Histology								
		I			II			III		
		1	2	3	1	2	3	1	2	3
0–2		9	12	42	5	4	28	1	1	19
		(157	134	212	77	71	130	21	22	101)
2–4		2	7	26	2	3	19	1	1	11
		(139	110	136	68	63	72	17	18	63)
4–6		9	5	12	3	5	10	1	3	7
		(126	96	90	63	58	42	14	14	43)
6–8		10	10	10	2	4	5	1	1	6
		(102	86	64	55	42	21	12	10	32)
8–10		1	4	5	2	2	0	0	0	3
		(88	66	47	50	35	14	10	8	21)
10–12		3	3	4	2	1	3	1	0	3
		(82	59	39	45	32	13	8	8	14)
12 +		1	4	1	2	4	2	0	2	3
		(76	51	29	42	28	7	6	6	10)

[a] Values in parentheses represent total follow-up months at risk.
Source: T. Holford, *Biometrics* **36**: 299–305 (1980), shown with permission of John Wiley & Sons, Inc.; the data are in the `Cancer` data file at the text website.

Let μ_{ijk} denote the expected number of deaths and t_{ijk} the total time at risk for histology i and stage of disease j, in follow-up time interval k. The Poisson loglinear model for the death rate,

$$\log(\mu_{ijk}/t_{ijk}) = \beta_0 + \beta_i^H + \beta_j^S + \beta_k^T,$$

treats each explanatory variable as a qualitative factor, where the superscript notation shows the classification labels. It has residual deviance 43.92 ($df = 52$). Here is some edited R output:

```
------------------------------------------------------------------------------
> Cancer <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Cancer.dat",
+                      header=TRUE)
> Cancer
   time histology stage count risktime
1     1         1     1     9      157 # 63 contingency table cells
...
63    7         3     3     3       10
> logrisktime = log(Cancer$risktime)
> fit <- glm(count ~ factor(histology) + factor(stage) + factor(time),
+            family=poisson, offset=logrisktime, data=Cancer)
> summary(fit)
                   Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)         -3.0093      0.1665   -18.073    <2e-16
factor(histology)2   0.1624      0.1219     1.332    0.1828
factor(histology)3   0.1075      0.1474     0.729    0.4658
factor(stage)2       0.4700      0.1744     2.694    0.0070
factor(stage)3       1.3243      0.1520     8.709    <2e-16
factor(time)2       -0.1274      0.1491    -0.855    0.3926 # showing 2 of 6
...                                                        # time effects
factor(time)7       -0.1752      0.2498    -0.701    0.4832
---
    Null deviance: 175.718  on 62  degrees of freedom
Residual deviance:  43.923  on 52  degrees of freedom

> library(car)
> Anova(fit) # likelihood-ratio tests of effects, adjusting for the others
                    LR Chisq  Df  Pr(>Chisq)
factor(histology)     1.876   2     0.39132
factor(stage)        99.155   2     < 2e-16
factor(time)         11.383   6     0.07724 .
------------------------------------------------------------------------------
```

The estimated stage-of-disease effects (0 for stage 1, 0.470 for stage 2, 1.324 for stage 3) show the progressively worsening death rate as the stage advances. The estimated death rate at the third stage of disease is $\exp(1.324) = 3.76$ times that at the first stage, adjusting for follow-up time and histology, with Wald 95% confidence interval $\exp[1.324 \pm 1.96(0.152)]$, or (2.79, 5.06). Likelihood-ratio tests indicate that, although stage of disease is an important prognostic factor, histology did not contribute significant additional information. Adding interaction terms between stage of disease and follow-up time interval does not significantly improve the fit (change in deviance $= 14.86$ with $df = 12$).

Models that assume a lack of interaction between follow-up time interval and either prognostic factor are called *proportional hazards* models. They have the same effects of histology and stage of disease in each time interval. Then, a ratio of hazards for two groups is the same at all times.

7.6.3 Negative Binomial Regression Models

Like the Poisson, the *negative binomial* is a probability distribution concentrated on the nonnegative integers. Unlike the Poisson, it has an additional parameter such that the variance can exceed the mean. It has a formula for probabilities, but we will only need to use its mean and variance,

$$E(Y) = \mu, \quad \text{var}(Y) = \mu + D\mu^2.$$

The index D, which is nonnegative, is called a *dispersion parameter*. The farther D falls above 0, the greater the overdispersion relative to Poisson variability. As D decreases toward 0, var(Y) decreases toward μ, and the negative binomial distribution converges to the Poisson distribution.

Negative binomial GLMs for counts or rates express μ in terms of explanatory variables. Most common is the $\log$ link, as in Poisson loglinear models, but sometimes the identity link is adequate. It is common to assume that the dispersion parameter D takes the same value at all values of the explanatory variables, much as ordinary regression models for a normal response take the variance parameter to be constant.

7.6.4 Example: Female Horseshoe Crab Satellites Revisited

In Section 3.3.3, we modeled $Y =$ number of male satellites for female horseshoe crabs. Using $x =$ width of the shell, the Poisson loglinear model has fit

$$\log(\hat{\mu}) = -3.30 + 0.164x$$

with $SE = 0.020$ for $\hat{\beta}$, as shown again in the following R output:

```
------------------------------------------------------------------------------
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
+                     header=TRUE)
> Crabs
    crab sat weight width color spine # sat is the count of satellites
1      1   8  3.050  28.3     2     3 # 173 lines in data file
...
173  173   0  2.000  24.5     2     2
> fit.pois <- glm(sat ~ width, family=poisson, data=Crabs)
> summary(fit.pois)
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -3.30476     0.54224   -6.095   1.1e-09
width         0.16405     0.01997    8.216   < 2e-16
---
Residual deviance: 567.88  on 171  degrees of freedom

> library(MASS)
> fit.negbin <- glm.nb(sat ~ width, data=Crabs) # negative binomial GLM
> summary(fit.negbin)
```

```
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -4.05251     1.17143   -3.459  0.000541
width         0.19207     0.04406    4.360   1.3e-05
---
(Dispersion parameter for Negative Binomial(0.9046) family taken to be 1)
              # D in our notation is 1/0.9046 = 1.11
Residual deviance: 195.81  on 171  degrees of freedom
--------------------------------------------------------------------------
```

The corresponding negative binomial GLM has

$$\log(\hat{\mu}) = -4.05 + 0.192x$$

with $SE = 0.044$ for $\hat{\beta}$. Moreover, $\hat{D} = 1.1$, so at an estimated mean $\hat{\mu}$, the estimated variance is $\hat{\mu} + 1.1\hat{\mu}^2$, compared to $\hat{\mu}$ for the Poisson GLM. Fitted values are similar, but the greater estimated variance in the negative binomial model and the resulting greater SE for $\hat{\beta}$ reflect the overdispersion uncaptured with the Poisson GLM. Inspection of Figure 3.3 shows that some zero counts occur even when the sample mean response is relatively large, reflecting this overdispersion.

For the Poisson model, the 95% Wald confidence interval for the effect of width (β) is $0.164 \pm 1.96(0.020)$, which is (0.125, 0.203). For the negative binomial model, it is $0.192 \pm 1.96(0.044)$, which is (0.105, 0.278). The profile likelihood confidence intervals are similar. Confidence intervals for β with the Poisson GLM are unrealistically narrow, because of not allowing for the overdispersion.

EXERCISES

7.1 For the recent General Social Survey data on $X =$ gender (males, females) and $Y =$ belief in an afterlife (no, yes), Table 7.12 shows results of fitting the independence loglinear model.

a. The deviance is 0.82 with $df = 1$. What does this suggest?

b. Report $\{\hat{\lambda}_j^Y\}$. Interpret $\hat{\lambda}_1^Y - \hat{\lambda}_2^Y$.

c. For the saturated model, software reports for $\{\hat{\lambda}_{ij}^{XY}\}$:

```
-------------------------------------------------------------------
                           Estimate  Std Error
genderfemales:beliefyes     0.1368     0.1507
-------------------------------------------------------------------
```

Estimate the odds ratio.

Table 7.12 Software output for Exercise 7.1 on belief in afterlife.

```
-------------------------------------------------------------------------------
                Estimate   Std. Error
 Intercept         4.5849       0.0752
 genderfemales     0.2192       0.0599
 beliefyes         1.4165       0.0752
-------------------------------------------------------------------------------
```

d. The text website has a data file `Postlife` that cross-classifies belief in life after death with race (black, white, other). Fit the independence model and interpret $\hat{\lambda}_1^Y - \hat{\lambda}_2^Y$.

7.2 For Table 2.9, let $D =$ defendant's race, $V =$ victims' race, and $P =$ death penalty verdict. Table 7.13 shows the output for fitting loglinear model (DV, DP, PV).

a. Report the estimated conditional odds ratio between D and P at each category of V. Interpret.

b. Test the goodness of fit of this model. Interpret.

c. Using the `DeathPenalty` data file at the text website, obtain the results shown in this output. Specify the corresponding logistic model with P as the response. Set up a grouped-data file to fit that model and show how estimated effects of D and V relate to loglinear model estimates. Which model seems more relevant for these data?

Table 7.13 Software output for Exercise 7.2 on death penalty verdicts.

```
-------------------------------------------------------------------------------
 Coefficients: # not showing intercept and main effect terms
               Estimate  Std. Error  z value  Pr(>|z|)
Dwhite:Vwhite   4.59497     0.31353   14.656   < 2e-16
Dwhite:Pyes    -0.86780     0.36707   -2.364    0.0181
Vwhite:Pyes     2.40444     0.60061    4.003  6.25e-05
---
Residual deviance:    0.37984  on 1  degrees of freedom
-------------------------------------------------------------------------------
```

7.3 Table 7.14 is based on automobile accident records supplied by the state of Florida Department of Highway Safety and Motor Vehicles. Subjects were classified by whether they were wearing a seat belt, whether ejected, and whether killed.

a. Find a loglinear model that describes the data well. Interpret the associations.

b. Since the sample size is large, goodness-of-fit statistics are large unless the model fits very well. Calculate the dissimilarity index for the model you found in (**a**) and interpret.

c. Conduct a Bayesian analysis of the homogeneous association model. Find and interpret posterior intervals for the conditional odds ratios. How do interpretations differ from ones you make with a frequentist analysis?

Table 7.14 Data for Exercise 7.3.

Safety Equipment in Use	Whether Ejected	Injury	
		Nonfatal	Fatal
Seat belt	Yes	1,105	14
	No	411,111	483
None	Yes	4,624	497
	No	157,342	1008

Source: Florida Department of Highway Safety and Motor Vehicles.

7.4 At the website `www.stat.ufl.edu/~aa/intro-cda/data` for the second edition of this book, the `MBTI` data file cross-classifies the MBTI Step II National Sample on four binary scales of the Myers–Briggs personality test: Extroversion/Introversion (E/I), Sensing/iNtuitive (S/N), Thinking/Feeling (T/F), and Judging/Perceiving (J/P). Fit the loglinear model of homogeneous association and conduct a goodness-of-fit test. Based on the fit, show that (i) the estimated conditional association is strongest between the S/N and J/P scales, (ii) there is not strong evidence of conditional association between the E/I and T/F scale or between the E/I and J/P scales.

7.5 Refer to the auto accident injury data shown in Table 7.5.

a. Explain why the fitted odds ratios in Table 7.7 for model (GL, GS, LS, GI, LI, SI) suggest that the most likely case for injury is accidents for females not wearing seat belts in rural locations.

b. Consider the following two-stage model. The first stage is a logistic model with S as the response, for the three-way $G \times L \times S$ table. The second stage is a logistic model with these three variables as predictors for I in the four-way table. Explain why this composite model is sensible, fit the models, and interpret results.

7.6 Table 7.15, from a General Social Survey, relates responses on $R =$ religious service attendance ($1 =$ at most a few times a year, $2 =$ at least several times a year), $P =$ political views ($1 =$ liberal, $2 =$ moderate, $3 =$ conservative), $B =$ birth control availability to teenagers between the ages of 14 and 16 ($1 =$ agree, $2 =$ disagree), $S =$ sex relations before marriage ($1 =$ wrong only sometimes or not wrong at all, $2 =$ always or almost always wrong).

a. Investigate the complexity needed for loglinear modeling by fitting models having only single-factor terms, all two-factor terms, and all three-factor terms. Select a model and interpret it by estimating conditional odds ratios.

b. Draw the independence graph for model (BP, BR, BS, PS, RS). Remark on conditional independence patterns. Are any fitted marginal and conditional associations identical?

c. Fit the loglinear model that corresponds to the logistic model that predicts S using the other variables as main effects, without any interaction. Does it fit adequately?

Table 7.15 Data (file `BPRS` at text website) for Exercise 7.6.

		Premarital Sex							
		1				2			
	Religious Attendence	1		2		1		2	
	Birth control	1	2	1	2	1	2	1	2
Political	1	99	15	73	25	8	4	24	22
Views	2	73	20	87	37	20	13	50	60
	3	51	19	51	36	6	12	33	88

7.7 The data in Table 7.16, from a General Social Survey, are in the `Spending` data file at the text website. Subjects were asked about government spending on the environment (E), health (H), assistance to big cities (C), and law enforcement (L).

Table 7.16 Opinions about government spending.

Cities										
Law Enforcement		1			2			3		
Environment	Health	1	2	3	1	2	3	1	2	3
1	1	62	17	5	90	42	3	74	31	11
	2	11	7	0	22	18	1	19	14	3
	3	2	3	1	2	0	1	1	3	1
2	1	11	3	0	21	13	2	20	8	3
	2	1	4	0	6	9	0	6	5	2
	3	1	0	1	2	1	1	4	3	1
3	1	3	0	0	2	1	0	9	2	1
	2	1	0	0	2	1	0	4	2	0
	3	1	0	0	0	0	0	1	2	3

The common response scale was (1 = too little, 2 = about right, 3 = too much). Compare the models with all two-factor and with all three-factor terms. For the homogeneous association model. estimate the conditional odds ratios using the *too much* and *too little* categories for each pair of variables. Summarize the associations. Based on these results, which term(s) might you consider dropping from the model? Why?

7.8 For a three-way contingency table, consider the independence graph,

$$X \text{————} Z \qquad\qquad Y$$

Write the corresponding loglinear model. Which pairs of variables are conditionally independent? Which pairs of variables have the same marginal association as their conditional association?

7.9 For a multiway contingency table, when is a logistic model more appropriate than a loglinear model? When is a loglinear model more appropriate?

7.10 For loglinear model (WXZ, WYZ), draw its independence graph and identify variables that are conditionally independent.

7.11 Refer to Exercise 7.7 with Table 7.16 and the `Spending` data file.

a. Beginning with the homogeneous association model, show that backward elimination yields (CE, CL, EH, HL). Interpret its fit.

b. Based on the independence graph for (CE, CL, EH, HL), show that (i) every path between C and H involves a variable in $\{E, L\}$; (ii) collapsing over H, one obtains the same associations between C and E and between C and L, and, collapsing over C, one obtains the same associations between H and E and between H and L; (iii) the conditional independence patterns between C and H and between E and L are not collapsible.

7.12 For the substance use data in Table 7.8, consider loglinear model $(AC, AM, CM, AG, AR, GM, GR)$.

a. Explain why the *AM* conditional odds ratio is unchanged by collapsing over race, but it is not unchanged by collapsing over gender.

b. Suppose we remove the *GM* term from the model. Construct the independence graph and show that $\{G, R\}$ are separated from $\{C, M\}$ by A. Explain why all conditional associations among A, C, and M are then identical to those in model (AC, AM, CM), collapsing over G and R.

7.13 Table 7.17 comes from a General Social Survey. Subjects were asked whether methods of birth control should be available to teenagers and how often they attend religious services.

a. Fit the independence model. Describe the lack of fit.

b. Using equally spaced scores, fit the linear-by-linear association model. Describe the association. Test goodness of fit. Test independence by using the ordinality, and interpret.

c. Fit the $L \times L$ model using column scores $\{1, 2, 4, 5\}$. Explain why a fitted local log odds ratio using columns 2 and 3 is double a fitted local log odds ratio using columns 1 and 2 or columns 3 and 4. What is the relation between the odds ratios?

Table 7.17 Data for Exercise 7.13 on religion and birth control.

	Teenage Birth Control			
Religious Attendance	Strongly Agree	Agree	Disagree	Strongly Disagree
Never	49	49	19	9
Less than once a year	31	27	11	11
Once or twice a year	46	55	25	8
Several times a year	34	37	19	7
About once a month	21	22	14	16
2–3 times a month	26	36	16	16
Nearly every week	8	16	15	11
Every week	32	65	57	61
Several times a week	4	17	16	20

7.14 True or false?

a. With a single categorical response variable, logistic regression models are more appropriate than loglinear models.

b. To model the association and interaction structure among several categorical response variables, logistic regression models are more appropriate than loglinear models.

c. The logistic model is a GLM assuming a binomial random component whereas the loglinear model is a GLM assuming a Poisson random component. Hence, when both are fitted to a contingency table having 50 cells with a binary response, the logistic model treats the cell counts as 25 binomial observations whereas the loglinear model treats the cell counts as 50 Poisson observations.

7.15 Consider Table 7.11 on survival of lung cancer patients.

a. Fit the more complex model that allows interaction between stage and time. Analyze whether it provides a significantly improved fit.

b. Fit the simpler model with main effects of stage and time and no histology effects. Check the fit of the model and interpret the estimated effects of stage of disease.

7.16 For the `Crabs` data file at the text website, let $y =$ number of satellites and $x =$ weight. Fit the Poisson and negative binomial (NB) loglinear models. For each model, report the prediction equation and SE of the weight effect, and construct a 95% confidence interval for β. Explain why the interval is wider with the NB model. Which model is more appropriate? Why?

7.17 A recent General Social Survey asked subjects how many times they had sexual intercourse in the previous month. The sample means were 5.9 for males and 4.3 for females; the sample variances were 54.8 and 34.4.

a. Does an ordinary Poisson GLM seem appropriate for comparing the means? Explain.

b. The GLM with log link and an indicator variable for gender ($1 =$ males, $0 =$ females) has gender estimate 0.308. The *SE* is 0.038 assuming a Poisson distribution and 0.127 assuming a negative binomial model. Why are the SE values so different?

c. The Wald 95% confidence interval for the ratio of means is $(1.26, 1.47)$ for the Poisson model and $(1.06, 1.75)$ for the negative binomial model. Which interval do you think is more appropriate? Why?

CHAPTER 8

MODELS FOR MATCHED PAIRS

This chapter introduces methods for comparing categorical responses for two samples that have a natural pairing between each subject in one sample and a subject in the other sample. The responses in the two samples are *matched pairs*. Because of the matching, the samples are statistically *dependent*. Methods that treat the two sets of observations as independent samples are inappropriate. Although the comparison estimates are fine, their standard errors are invalid when observations are actually correlated.

The most common way that dependent samples occur is when each sample has the same subjects. This happens in longitudinal studies, which observe the same subjects at several times. It also happens in surveys that observe two or more similar response variables that have the same categories. Table 8.1 illustrates such a case, for data from a General Social Survey. Subjects were asked whether, to help the environment, they would be willing to (1) pay higher taxes, (2) accept a cut in living standards. The rows of the table are the categories for opinion about paying higher taxes. The columns are the same categories for opinion about accepting a cut in living standards. The marginal counts display the outcome frequencies for the two samples. The row marginal counts (359, 785) are the (yes, no) totals for paying higher taxes. The column marginal counts (334, 810) are the (yes, no) totals for accepting a cut in living standards.

After presenting inferential methods for comparing dependent proportions, we introduce logistic regression models for such data. We then extend the methods to multicategory responses, both nominal and ordinal scale. Finally, we present two applications that yield

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

Table 8.1 Opinions relating to environment, with unknown cell probabilities in parentheses.

Pay Higher Taxes	Cut Living Standards		
	Yes	No	Total
Yes	227 (π_{11})	132 (π_{12})	359
No	107 (π_{21})	678 (π_{22})	785
Total	334	810	1144

categorical matched-pairs data — measuring agreement between two observers who each evaluate the same subjects on a categorical scale and summarizing pairwise comparisons of categories that result in a preference for one category over another.

8.1 COMPARING DEPENDENT PROPORTIONS FOR BINARY MATCHED PAIRS

For Table 8.1, how can we compare the probabilities of a *yes* outcome for the two environmental questions? Of the 1144 subjects cross-classified in this table, the sample proportions of subjects who said *yes* were $359/1144 = 0.314$ for raising taxes and $334/1144 = 0.292$ for accepting cuts in living standards. The same people responded to each question, so these marginal proportions are correlated. A strong association exists between opinions on the two questions, the sample odds ratio being $(227 \times 678)/(132 \times 107) = 10.9$.

Let (y_1, y_2) denote the row and column responses for the matched pairs. Let n_{ij} denote the number of subjects who make response i for y_1 and response j for y_2. We can summarize the counts in a 2×2 contingency table, Table 8.1 being an example. Let $\{\pi_{ij}\}$ denote the corresponding cell probabilities, also shown in Table 8.1. The probability of a *yes* outcome is $P(Y_1 = 1) = \pi_{11} + \pi_{12}$ for question 1 and $P(Y_2 = 1) = \pi_{11} + \pi_{21}$ for question 2. When these are identical, the probabilities of a *no* outcome are also identical. There is then said to be *marginal homogeneity*. Since

$$P(Y_1 = 1) - P(Y_2 = 1) = (\pi_{11} + \pi_{12}) - (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21},$$

marginal homogeneity in 2×2 tables is equivalent to $\pi_{12} = \pi_{21}$.

8.1.1 McNemar Test Comparing Marginal Proportions

For matched-pairs data with a binary response, a test of marginal homogeneity has null hypothesis

$$H_0 : P(Y_1 = 1) = P(Y_2 = 1), \text{ or equivalently } H_0 : \pi_{12} = \pi_{21}.$$

When H_0 is true, we expect similar values for n_{12} and n_{21}. Let $n^* = n_{12} + n_{21}$ denote the total count in these two cells. Their allocations to those cells are outcomes of a binomial

variate. Under H_0, each of these n^* observations has a $\frac{1}{2}$ chance of contributing to n_{12} and a $\frac{1}{2}$ chance of contributing to n_{21}. Therefore, n_{12} and n_{21} are numbers of *successes* and *failures* for a binomial distribution having n^* trials and success probability $\frac{1}{2}$.

When $n^* > 10$, the binomial distribution has a similar shape to the normal distribution with the same mean, which is $\frac{1}{2}n^*$, and standard deviation, which is $\sqrt{n^*(\frac{1}{2})(\frac{1}{2})}$. The standardized normal test statistic equals

$$z = \frac{n_{12} - (\frac{1}{2})n^*}{\sqrt{n^*(\frac{1}{2})(\frac{1}{2})}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}. \tag{8.1}$$

For Table 8.1, $z = (132 - 107)/\sqrt{132 + 107} = 1.62$. The two-sided P-value is 0.106. The evidence against marginal homogeneity is weak.

Most software, such as the following R output, reports the square of the z statistic, which has an approximate chi-squared distribution with $df = 1$. For these data, it takes value $(1.62)^2 = 2.62$. This test for a comparison of two dependent proportions is called[1] the *McNemar test*. Here is some R output for this test and follow-up confidence intervals:

```
------------------------------------------------------------------------------
> Opinions <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                          Envir_opinions.dat", header=TRUE)
> Opinions
      person y1 y2 # Envir_opinions data file is at text website
1          1  1  1 # Data file has 1144 lines, one for each of 1144 people
2          2  1  1 # Each person has two binary responses (y1 and y2)
...
1144   1144  2  2
> tab <- xtabs(~y1 + y2, data = Opinions)
> tab
   y2
y1    1   2
  1 227 132
  2 107 678
> mcnemar.test(tab, correct=FALSE) # don't use continuity correction,
        McNemar's Chi-squared test # which is too conservative
McNemar's chi-squared = 2.6151, df = 1, p-value = 0.1059
> library(PropCIs)
> diffpropci.Wald.mp(107, 132, 1144, 0.95) # (n21, n12, n, conf. level)
> # or, diffpropci.Wald.mp(tab[2, 1], tab[1, 2], sum(tab), 0.95)
 -0.00460  0.04831 # 95% Wald CI for difference of marginal probabilities
> scoreci.mp(tab[2, 1], tab[1, 2], sum(tab), 0.95)
 -0.00466  0.04849 # 95% score CI for difference of marginal probabilities
------------------------------------------------------------------------------
```

[1] Named after the psychologist–statistician who proposed the test in 1947.

8.1.2 Estimating the Difference between Dependent Proportions

A confidence interval for $P(Y_1 = 1) - P(Y_2 = 1)$ is more informative than a significance test. In terms of the cell counts that McNemar's test uses and the overall sample size n, the standard error of the sample difference of marginal proportions is

$$SE = \frac{1}{n}\sqrt{(n_{12} + n_{21}) - (n_{12} - n_{21})^2/n}.$$

The Wald confidence interval adds and subtracts the relevant z-score times this standard error to obtain a margin of error.

For small n^* or when $P(Y_1 = 1)$ and $P(Y_1 = 1)$ are both near 0 or 1, the Wald interval has actual coverage probability well below the nominal level. A simple fix that improves its performance is to add 2 observations before applying the Wald formula, by adding 0.5 to every cell of the 2×2 table.[2] Also better is the interval based on correspondence with the score test, which is computationally complex but available in R as shown in the above output.

For Table 8.1, the difference of sample marginal proportions is $0.314 - 0.292 = 0.022$. For $n = 1144$ observations with $n_{12} = 132$ and $n_{21} = 107$, we obtain $SE = 0.0135$, so the Wald 95% confidence interval equals $0.022 \pm 1.96(0.0135)$, or $(-0.005, 0.048)$. This is also the score confidence interval. We infer that the probability of a *yes* response was between 0.005 less and 0.048 higher for paying higher taxes than for accepting a cut in living standards. If the probabilities differ, the difference is apparently small.

8.2 MARGINAL MODELS AND SUBJECT-SPECIFIC MODELS FOR MATCHED PAIRS

Models for binary data, such as logistic regression, extend to handle matched-pairs responses. In fact, the analyses of the previous section also occur as by-products of model-fitting.

8.2.1 Marginal Models for Marginal Proportions

The difference between the marginal probabilities occurs as a parameter in a model using the identity link function. For the model

$$P(Y_1 = 1) = \alpha + \delta, \;\; P(Y_2 = 1) = \alpha,$$

the parameter $\delta = P(Y_1 = 1) - P(Y_2 = 1)$. This is equivalent to

$$P(Y_t = 1) = \alpha + \delta x_t,$$

[2] A. Agresti and Y. Min, *Statistics in Medicine* **24**: 729–740 (2005), available with the `diffpropci.mp` function in the `PropCIs` package in R.

where x_t is an indicator variable that equals 1 when $t = 1$ and 0 when $t = 2$. The hypothesis of marginal homogeneity is H_0: $\delta = 0$. An alternative model applies the logit link,

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta x_t. \tag{8.2}$$

Its parameter β is the log odds ratio comparing the marginal distributions.

These models are called *marginal models*. They focus on the marginal distributions of the two responses. The next chapter presents a method, called *generalized estimating equations* (GEE), used for fitting marginal models.

8.2.2 Example: Environmental Opinions Revisited

For Table 8.1 with opinions about Y_1 = raising taxes and Y_2 = accepting cuts in living standards, the following R output uses the GEE method to fit the two marginal models just mentioned. Here, the data file has a different format than in the previous R output, a format explained in Section 8.2.3. It has indicator variables for the question (1 for question 1 about taxes, 0 for question 2 about living standards) and for each response variable (1 for *yes*, 0 for *no*):

```
-------------------------------------------------------------------------------
> Opinions <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Opinions.dat",
+                        header=TRUE)
> Opinions # data file at text website has 2 lines for each person
     person question y # y variable is y1 when question=1, y2 when question=0
1         1        1 1 # y1 for person 1
2         1        0 1 # y2 for person 1
3         2        1 1
4         2        0 1
...
2287   1144        1 0 # y1 for person 1144
2288   1144        0 0 # y2 for person 1144
> library(gee)
> fit <- gee(y ~ question, id=person, family=binomial(link="identity"),
+            data=Opinions) # id identifies variable on which observe y1, y2
> summary(fit) # question para. for identity link is difference of proportions
             Estimate  Naive S.E.   Naive z  Robust S.E.  Robust z
(Intercept)   0.29196     0.01345  21.70970      0.01344  21.71920
question      0.02185     0.01922   1.13725      0.01350   1.61897

> fit2 <- gee(y ~ question, id=person, family=binomial(link=logit),
+             data=Opinions)
> summary(fit2) # question parameter for logit link is log odds ratio
             Estimate  Naive S.E.    Naive z  Robust S.E.   Robust z
(Intercept)  -0.88589     0.06506  -13.61740      0.06503  -13.62336
question      0.10353     0.09108    1.13674      0.06398    1.61824
-------------------------------------------------------------------------------
```

In the GEE output, the "naive" standard errors treat the two observations for each person as independent, which is inappropriate, whereas the "robust" standard errors account for the dependence. Matched-pairs data usually exhibit a strong positive association between Y_1 and Y_2. When the sample odds ratio exceeds 1.0 (and it is 10.9 for these data), the robust SE for the estimate of the effect parameter is smaller than the SE that naively assumes independent samples. An advantage of using dependent samples, compared with independent samples, is a more precise estimate of the effect, such as the difference of proportions.

For the model with an identity link, the estimated difference of proportions is $\hat{\delta} = 0.314 - 0.292 = 0.022$. In Section 8.1.2 we found the estimate 0.022 and robust $SE = 0.0135$ shown in the output. For the model with a logit link, $\hat{\beta}$ is the log odds ratio for the sample marginal distributions, $\hat{\beta} = \log[(359/785)/(334/810)] = 0.104$. The odds ratio estimate is $\exp(\hat{\beta}) = 1.11$. The population odds of willingness to pay higher taxes are estimated to be 11% higher than the population odds of willingness to accept cuts in living standards.

The `Opinions` data file at the text website that we used in these `R` marginal model analyses has a form that we discuss next. Rather than showing the two responses y_1 and y_2 for a person on a single line (as in the data file for the output in Section 8.1.1), it shows the responses on two separate lines, one line for y_1 and one line for y_2.

8.2.3 Subject-Specific and Population-Averaged Tables

An alternate representation for binary matched-pairs presents the data in n separate 2×2 partial tables, one for each matched pair. Partial table i shows the responses (y_{i1}, y_{i2}) for the ith matched pair. It has columns that are the two possible outcomes for each observation. It shows y_{i1} in row 1 and y_{i2} in row 2.

Table 8.1 cross-classified results on two environmental questions for 1144 subjects. Table 8.2 shows the partial table for a person who answered *yes* on both questions.

Table 8.2 Representation of subject-specific table for matched pair contributing to count $n_{11} = 227$ in Table 8.1.

	Response	
Question	Yes	No
y_{i1}: Pay higher taxes?	1	0
y_{i2}: Cut living standards?	1	0

Each of the 1144 subjects has a partial table, displaying the two matched observations; 227 look like Table 8.2, 132 have first row (1, 0) and second row (0, 1), representing *yes* on question 1 and *no* on question 2, 107 have first row (0, 1) and second row (1, 0), and 678 have (0, 1) in each row. The 1144 subjects provide 2288 observations in a $2 \times 2 \times 1144$ contingency table. The data file in the `R` output in Section 8.2.2 has this form, but shows only the first column (for the *yes* outcome), because the second column is redundant (e.g., outcome 1 for *yes* implies outcome 0 for *no*).

For binary matched-pairs data, we refer to the $2 \times 2 \times n$ table with a separate 2×2 partial table for each of n matched pairs as the *subject-specific table*. By contrast, the 2×2 table that cross-classifies in a single table the two responses for all n subjects is called

a *population-averaged table*. Table 8.1 is an example. Its margins provide estimates of population marginal probabilities. In fact, if we collapse the $2 \times 2 \times 1144$ subject-specific table over the 1144 subjects, we obtain a 2×2 table with first row equal to (359, 785) and the second row equal to (334, 810), which are the marginal counts in Table 8.1.

8.2.4 Conditional Logistic Regression for Matched-Pairs *

An alternative model for matched-pairs data permits each subject to have their own probability distribution, by having a separate intercept term for each subject. With logit link, the model is

$$\text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta x_{it}, \;\; t = 1, 2, \tag{8.3}$$

where x_{it} is an indicator variable for subject i that equals 1 when $t = 1$ and 0 when $t = 2$. The probabilities of the *success* outcome for subject i equal

$$P(Y_{i1} = 1) = \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)}, \qquad P(Y_{i2} = 1) = \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)}.$$

A subject with a relatively large positive α_i (compared to the magnitude of β) has a high probability of success for each observation and is likely to have a success for each. A subject with a relatively large negative α_i has a low probability of success for each observation and is likely to have a failure for each. The greater the variability in values of these parameters, the greater the overall positive association between the two observations, successes (failures) for y_{i1} tending to occur with successes (failures) for y_{i2}.

Such a model naturally refers to the subject-specific partial tables of form Table 8.2. Because the model has a term for each subject, the β effect comparing the responses is *conditional* on the subject. The model assumes that, for each subject, the odds of success for y_{i1} are $\exp(\beta)$ times the odds of success for y_{i2}. Since each partial table refers to a single subject, this conditional association is a *subject-specific effect*. When $\beta = 0$, for each subject, the probability of success is the same for both observations. This implies marginal homogeneity.

Inference for this model focuses on β for comparing the distributions. The $\{\alpha_i\}$ subject parameters permit heterogeneity among subjects, but their values are not usually of interest. In fact, by permitting a separate intercept for each subject, the model has as many $\{\alpha_i\}$ parameters as subjects. Therefore, the number of parameters grows as the sample size n grows. This causes difficulties with the fitting process and with the properties of ordinary ML estimators. For example, if we try to estimate all these parameters at once, the ML estimator $\hat{\beta}$ actually is badly biased, actually approaching 2β as n grows. One remedy, *conditional logistic regression*, maximizes the likelihood function and finds $\hat{\beta}$ for a conditional distribution that eliminates the $\{\alpha_i\}$. (Section 5.4.5 introduced this method for conducting inference for logistic regression.) For tables with counts $\{n_{ij}\}$ summarizing the cross-classification of y_1 and y_2, such as Table 8.1, the conditional ML estimate of the odds ratio $\exp(\beta)$ for model (8.3) equals n_{12}/n_{21}. For Table 8.1, $\exp(\hat{\beta}) = 132/107 = 1.23$. Assuming the model holds, a subject's estimated odds of a *yes* response are 23% higher for raising taxes (question 1) than for accepting a lower standard of living (question 2).

By contrast, the odds ratio of 1.11 found above in Section 8.2.1 refers to the margins of Table 8.1, which equivalently are the rows of the marginal table obtained by collapsing the $2 \times 2 \times 1144$ subject-specific contingency table. That these odds ratios take different values merely reflects how conditional (subject-specific) odds ratios in partial tables of a three-way contingency table can differ from odds ratios in marginal tables that collapse over the third variable.

As in the McNemar test, n_{12} and n_{21} provide all the information needed for inference about β for logistic model (8.3). An alternative way of fitting model (8.3), which Chapter 10 presents, treats $\{\alpha_i\}$ as *random effects*. This approach treats $\{\alpha_i\}$ as unobserved random variables having a normal distribution. In most cases, the ML estimate of β is then the same as with the conditional ML approach.

8.2.5 Logistic Regression for Matched Case-Control Studies *

Case-control studies[3] that match a single control with each case are an important application having matched-pairs data. For a binary response, each case ($y_{i1} = 1$) is matched with a control ($y_{i2} = 0$) according to certain criteria that could affect the response. The study observes the explanatory variable X for cases and controls and analyzes the *XY* association.

Table 8.3 shows results from a matched case-control study. A study of effects on birthweight matched each case in which the child was underweight with a control in which the child had normal weight. The mothers, who were matched according to their age, were asked whether they were smokers ($x = 0$, no; $x = 1$, yes). Table 8.3 has the same form as Table 8.1, except that the categories of X rather than the categories of Y form the two rows and the two columns.

Table 8.3 Smoking behavior for birthweight case-control pairs.

Normal Birth Weight (Controls)	Low Birth Weight (Cases)	
	Nonsmokers	Smokers
Nonsmokers	159	22
Smoker	8	14

Source: Partly based on data in B. Mukherjee, I. Liu, and S. Sinha, *Statist. Medic.* **26**: 3240–3257 (2007).

A display of the data using a partial table (similar to Table 8.2) for each matched case-control pair reverses the roles of X and Y. In matched pair i, one subject has $y_{i1} = 1$ (the case) and one subject has $y_{i2} = 0$ (the control). Table 8.4 shows the four possible patterns of x values. There are 159 partial tables of type 8.4a, since for 159 pairs both the case and the control were nonsmokers, 22 partial tables of type 8.4b, 8 of type 8.4c, and 14 of type 8.4d.

For the logistic model (8.3), the odds that a smoker ($x = 1$) is a case equal $\exp(\beta)$ times the odds that a nonsmoker ($x = 0$) is a case. The probabilities in the model refer to the distribution of Y given X, but these retrospective data provide information only about the

[3] Sections 2.3.5 and 4.1.4 introduced the concept of case-control studies.

Table 8.4 Possible case-control pairs for Table 8.3.

	a		b		c		d	
Smoker	Case	Control	Case	Control	Case	Control	Case	Control
No	1	1	0	1	1	0	0	0
Yes	0	0	1	0	0	1	1	1

distribution of X given Y. We can estimate $\exp(\beta)$, however, since it refers to the *XY* odds ratio, which relates to both types of conditional distribution (Section 2.3.5). Even though a case-control study reverses the roles of X and Y in terms of which is fixed and which is random, the conditional ML estimate of the odds ratio $\exp(\beta)$ for Table 8.3 is $n_{12}/n_{21} = 22/8 = 2.75$.

For further details and examples of logistic regression in the context of case-control studies, see Hosmer et al. (2013, Chapter 7).

8.3 COMPARING PROPORTIONS FOR NOMINAL MATCHED-PAIRS RESPONSES

Categorical matched-pairs analyses generalize to $c > 2$ outcome categories. A square $c \times c$ contingency table $\{n_{ij}\}$ shows counts of possible outcomes (i, j) for (y_1, y_2). Marginal homogeneity is

$$P(Y_1 = i) = P(Y_2 = i) \quad \text{for } i = 1, \ldots, c,$$

under which each row marginal probability equals the corresponding column marginal probability.

8.3.1 Marginal Homogeneity for Baseline-Category Logit Models

For nominal-scale responses, Section 6.1 showed that standard models use baseline-category logits. For matched pairs, the relevant marginal model is

$$\log\left[\frac{P(Y_1 = j)}{P(Y_1 = c)}\right] = \alpha_j + \beta_j, \;\; \log\left[\frac{P(Y_2 = j)}{P(Y_2 = c)}\right] = \alpha_j,$$

for $j = 1, \ldots, c-1$.

The hypothesis of marginal homogeneity is H_0: $\beta_1 = \cdots \; \beta_{c-1} = 0$. A test of H_0 has $df = c - 1$. Software is not well developed for ML fitting of marginal models with baseline-category logits, but we can test H_0 using the generalized estimating equations (GEE) approach presented in the next chapter.

8.3.2 Example: Coffee Brand Market Share

A survey recorded the brand choice for a sample of buyers of instant coffee. At a later coffee purchase by these subjects, the brand choice was again recorded. Table 8.5 shows

Table 8.5 Choice of decaffeinated coffee at two purchase dates.

First Purchase	Second Purchase					
	High Pt	Taster's	Sanka	Nescafe	Brim	Total
High Point	93	17	44	7	10	171
Taster's Choice	9	46	11	0	9	75
Sanka	17	11	155	9	12	204
Nescafe	6	4	9	15	2	36
Brim	10	4	12	2	27	55
Total	135	82	231	33	60	541

Source: Based on data from R. Grover and V. Srinivasan, *J. Marketing Res.* **24**: 139–153 (1987); reprinted with permission by the American Marketing Association.

results for five brands of decaffeinated coffee. The relatively large cell counts on the main diagonal indicate that most buyers did not change their brand choice. The sample marginal proportions for the brands were (0.32, 0.14, 0.38, 0.07, 0.10) for the first purchase and (0.25, 0.15. 0.43, 0.06, 0.11) for the second purchase.

Using GEE methodology, we can fit the baseline-category logit model, as shown in the following R output:

```
-------------------------------------------------------------------------------
> Coffee <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Coffee.dat",
+                      header=TRUE)
> Coffee # subject-specific data file from text website
     person purchase y # purchase is 1 for first purchase, 0 for second
1         1        1 1
2         1        0 1
3         2        1 1
4         2        0 1
...
1081    541        1 5
1082    541        0 5
> library(multgee) # package for multinomial GEE analyses
> fit <- nomLORgee(y ~ purchase, id=person, LORstr="independence", data=Coffee)
> summary(fit) # nomLORgee uses baseline category logits for nominal response
Link : Baseline Category Logit
            Estimate   san.se    san.z  Pr(>|san.z|) # san.se = robust SE
beta01       0.81093  0.15516   5.2265       < 2e-16 # alpha1 in our notation
purchase:1   0.32340  0.16830   1.9215       0.05466
beta02       0.31237  0.16989   1.8387       0.06596 # alpha2 in our notation
purchase:2  -0.00222  0.18662  -0.0119       0.99051
beta03       1.34807  0.14490   9.3035       < 2e-16 # alpha3 in our notation
purchase:3  -0.03729  0.15807  -0.2359       0.81353
beta04      -0.59784  0.21672  -2.7585       0.00581 # alpha4 in our notation
purchase:4   0.17402  0.23531   0.7396       0.45957
---
> fit0 <- nomLORgee(y ~ 1, id=person, LORstr="independence", data=Coffee)
> waldts(fit0, fit)
```

```
Model under H_0: y ~ 1 # null model (marginal homogeneity)
Model under H_1: y ~ purchase
Wald Statistic = 12.4869, df=4, p-value=0.0141

> with(Coffee, mantelhaen.test(purchase, y, person))
Cochran-Mantel-Haenszel M^2 = 12.291, df = 4, p-value = 0.01531
--------------------------------------------------------------------------------
```

For instance, the first two parameter estimates refer to the logits for categories 1 and 5, using each margin. The estimated effect of $\hat{\beta}_1 = 0.323$ is the marginal log odds ratio for the 2×2 table with elements (171, 55) from row totals 1 and 5 and (135, 60) from column totals 1 and 5. To test H_0: $\beta_1 = \beta_2 = \beta_2 = \beta_4 = 0$, we fit the null model that satisfies this. The Wald statistic comparing the models is 12.49, based on $df = 4$, which provides evidence of marginal heterogeneity ($P = 0.014$).

To estimate the change for a given brand, we can combine the other categories and use the methods of Section 8.1. We illustrate by comparing the proportions selecting High Point at the two times. We construct the table with row and column categories (High Point, Others). This table has counts, by row, of (93, 78 / 42, 328). The z statistic (8.1) equals $(78 - 42)/\sqrt{78 + 42} = 3.3$, strong evidence of a change in the population proportion choosing High Point ($P = 0.001$). The estimated difference is $0.32 - 0.25 = 0.07$ and a 95% confidence interval is 0.07 ± 0.04. The small P-value for the overall test of marginal homogeneity mainly reflects a decrease in the proportion choosing High Point and an increase in the proportion choosing Sanka, with no evidence of change for the other brands.

8.3.3 Using the Cochran–Mantel–Haenszel Test to Test Marginal Homogeneity *

An alternative test of marginal homogeneity is a score test.[4] Its formula is beyond our scope, but for $c = 2$, it simplifies to the McNemar statistic, the square of (8.1).

The statistic for general c also has an approximate chi-squared distribution with $df = c - 1$. It is a special case of a statistic applied to the subject-specific data file called by software the *Cochran–Mantel–Haenszel* statistic,[5] as the statistic generalizes one proposed by these authors in the 1950s. For these data, we obtained test statistic 12.29 ($df = 4$), with P-value $= 0.015$, as also shown in the above R output.

8.3.4 Symmetry and Quasi-Symmetry Models for Square Contingency Tables *

A quite different approach to comparing marginal probabilities in square contingency tables is based on models for the joint distribution. The joint probabilities $\{\pi_{ij}\}$ in a square contingency table satisfy *symmetry* if

$$\pi_{ij} = \pi_{ji}$$

[4] This tests H_0: all $\beta_i = 0$ for the quasi-symmetry model of Section 8.3.4.

[5] For example, with the *mantelhaen.test* function in R or the *cmh* option in SAS (PROC FREQ).

for all pairs of cells. Cell probabilities on one side of the "main diagonal" (where $i = j$ and the row variable and column variable outcomes are the same) are a mirror image of those on the other side. When symmetry holds, necessarily marginal homogeneity also holds. When $c > 2$, though, marginal homogeneity can occur without symmetry. The symmetry condition has the simple logistic form

$$\log(\pi_{ij}/\pi_{ji}) = 0 \text{ for all } i \text{ and } j.$$

The standardized residuals for the symmetry model equal

$$r_{ij} = (n_{ij} - n_{ji})/\sqrt{n_{ij} + n_{ji}}.$$

For $c = 2$, this is the z statistic (8.1) for which z^2 is the McNemar statistic.

When the marginal distributions differ substantially, the symmetry model fits poorly. A generalized model that can accommodate marginal heterogeneity is the *quasi-symmetry* model,[6]

$$\log(\pi_{ij}/\pi_{ji}) = \beta_i - \beta_j \text{ for all } i \text{ and } j. \tag{8.4}$$

One parameter is redundant, and we set $\beta_1 = 0$ or $\beta_c = 0$. The higher the value of $\hat{\beta}_i - \hat{\beta}_j$, relatively more observations fall in the cell in row i and column j than in the cell in row j and column i.

To fit the quasi-symmetry model, you can treat each pair of cell counts (n_{ij}, n_{ji}) as an independent binomial variate. Set up c artificial explanatory variables, corresponding to the coefficients of the $\{\beta_i\}$ parameters. For the logit $\log(\pi_{ij}/\pi_{ji})$ for a given pair of categories, the variable for β_i is 1, the variable for β_j is -1, and the variables for the other parameters equal 0, as shown in the R code for the following example. The model has intercept forced to equal 0. The fit has marginal totals equal to the observed marginal totals.

For the quasi-symmetry model, marginal homogeneity is the special case in which all $\beta_i = 0$. However, this special case is the symmetry model. In other words, for the quasi-symmetry model, marginal homogeneity is equivalent to symmetry. To test H_0: marginal homogeneity, we can test the null hypothesis of symmetry against the alternative hypothesis of quasi-symmetry. The likelihood-ratio test compares the residual deviances for the two models. For $c \times c$ contingency tables, the test has $df = c - 1$.

8.3.5 Example: Coffee Brand Market Share Revisited *

We next fit the symmetry and quasi-symmetry model to the coffee brand data of Table 8.5. From the next R output, the symmetry model has deviance = 22.47, with $df = 10$. The lack of fit results primarily from the discrepancy between n_{13} and n_{31}. For that pair, the standardized residual equals $(44 - 17)/\sqrt{44 + 17} = 3.46$. Consumers of High Point changed to Sanka more often than the reverse. Otherwise, the symmetry model fits most of the table fairly well. The quasi-symmetry model has deviance = 9.97, with $df = 6$. Permitting the marginal distributions to differ yields a better fit. Here, $\hat{\beta}_1 - \hat{\beta}_3 = 0.595 - (-0.113) = 0.709$ means that the estimated probability of changing from High

[6] Proposed by Henri Caussinus, *Ann. Fac. Sci. Univ. Toulouse* **29**: 77–182 (1965).

Point to Sanka were $\exp(0.709) = 2.03$ times the estimated probability of changing from Sanka to High Point. Here is R code showing an appropriate data file for fitting the models:

```
------------------------------------------------------------------------------
> Coffee2 <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Coffee2.dat",
+                       header=TRUE)
> Coffee2 # Coffee2 data file for quasi-symmetry model at text website
   H  T  S  N  B nij nji # nij and nji are opposite cell counts in Table 8.5
1  1 -1  0  0  0  17   9 # counts for cells with H and T
2  1  0 -1  0  0  44  17 # counts for cells with H and S,
3  1  0  0 -1  0   7   6 # counts for cells with H and N, etc.
4  1  0  0  0 -1  10  10
5  0  1 -1  0  0  11  11
6  0  1  0 -1  0   0   4
7  0  1  0  0 -1   9   4
8  0  0  1 -1  0   9   9
9  0  0  1  0 -1  12  12
10 0  0  0  1 -1   2   2 # 10 pairs of opposite cells for nij and nji
> symm <- glm(nij/(nij+nji) ~ -1, family=binomial, weights=nij+nji,
+             data=Coffee2)
> summary(symm) # symmetry model; -1 in model statement sets intercept = 0
Residual deviance: 22.473  on 10  degrees of freedom

> QS <- glm(nij/(nij+nji) ~ -1 + H + T + S + N + B, family=binomial,
+           weights=nij+nji, data=Coffee2)
> summary(QS) # quasi-symmetry model
    Estimate  Std. Error  z value  Pr(>|z|)
H    0.59544     0.29366    2.028    0.0426
T   -0.00400     0.32936   -0.012    0.9903
S   -0.11330     0.28508   -0.397    0.6911
N    0.30212     0.40159    0.752    0.4519
B         NA          NA       NA        NA # one category has estimate = 0
---
   Null deviance: 22.473  on 10  degrees of freedom # symmetry
Residual deviance:  9.974  on  6  degrees of freedom # quasi symmetry
------------------------------------------------------------------------------
```

The deviance difference of $22.47 - 9.97 = 12.50$ between the symmetry model and the quasi-symmetry model, based on $df = 4$, provides evidence of marginal heterogeneity ($P = 0.014$).

8.4 COMPARING PROPORTIONS FOR ORDINAL MATCHED-PAIRS RESPONSES

The tests of marginal homogeneity for a nominal-scale response variable, having $df = c - 1$, are designed to detect *any* difference between the margins. They treat the categories as unordered, using all $c - 1$ degrees of freedom available for comparisons of c pairs of marginal proportions.

8.4.1 Marginal Homogeneity and Cumulative Logit Marginal Model

When the categories are ordered, tests can analyze whether responses tend to be higher in one margin than the other. Ordinal tests, which have $df = 1$, are usually much more powerful. This is especially true when c is large and the association between classifications is strong.

An ordinal model comparison of the margins can use logits of cumulative probabilities (Section 6.2). The marginal model

$$\text{logit}[P(Y_1 \leq j)] = \alpha_j + \beta, \;\; \text{logit}[P(Y_2 \leq j)] = \alpha_j$$

generalizes the marginal logit model (8.2). Like the cumulative logit models of Section 6.2, it has the *proportional odds* structure by which the effect β is the same for each cumulative probability. The model states that the odds that a random observation on Y_1 falls in category j or below, instead of above category j, are $\exp(\beta)$ times the odds for a random observation on Y_2.

8.4.2 Example: Recycle or Drive Less to Help the Environment?

Table 8.6 is from a General Social Survey. Subjects were asked "How often do you cut back on driving a car for environmental reasons?" and "How often do you make a special effort to sort glass or cans or plastic or papers and so on for recycling?"

Table 8.6 Behaviors on recycling and driving less to help the environment.

	Drive Less			
Recycle	Always	Often	Sometimes	Never
Always	12	43	163	233
Often	4	21	99	185
Sometimes	4	8	77	230
Never	0	1	18	132

Using GEE methodology introduced in the next chapter for marginal models, we obtain $\hat{\beta} = 2.754$ ($SE = 0.081$) for this cumulative logit model, as shown in the following R output:

```
------------------------------------------------------------------------------
> Envir <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Envir.dat",
+                     header=TRUE)
> Envir # subject-specific data file from text website
     person question y
1         1        1 1 # observation on each question for each of 1230 persons
2         1        0 1
...
2459   1230        1 4
2460   1230        0 4
> library(multgee) # package for multinomial GEE analyses
> fit <- ordLORgee(y ~ question, id=person, LORstr="independence", data=Envir)
> summary(fit) # ordLORgee uses cumulative logits for ordinal response
Link : Cumulative logit
```

```
         Estimate   san.se     san.z  Pr(>|san.z|) # san.se = robust SE
beta01   -3.35111  0.08289  -40.4287     < 2.2e-16 # alpha_1 in our notation
beta02   -2.27673  0.07430  -30.6424     < 2.2e-16 # alpha_2 (3 intercepts)
beta03   -0.58488  0.05882   -9.9443     < 2.2e-16 # alpha_3
question  2.75361  0.08147   33.7985     < 2.2e-16 # beta cumul. log odds ratio
------------------------------------------------------------------------------
```

The estimates[7] are the same as if we used ML and treated the pairs of responses as independent rather than correlated, but the standard errors are robust and incorporate the correlation. From the estimated cumulative odds ratio of $\exp(\hat{\beta}) = 15.7$, the estimated odds of response *always* on recycling, instead of the other three categories, are 15.7 times the estimated odds of response *always* for driving less. The effect is substantial. For H_0: $\beta = 0$, the Wald test statistic $z = \hat{\beta}/SE = 2.754/0.081 = 33.8$ provides extremely strong evidence against H_0: marginal homogeneity.

8.4.3 An Ordinal Quasi-Symmetry Model *

Alternative ways exist of using ordinality to compare margins of square contingency tables. This section presents a test using an ordinal logit model that generalizes the symmetry model and Section 10.3 introduces subject-specific cumulative logit models.

In Section 8.3.4 we used the quasi-symmetry model to compare margins of a square table. However, that model treats the classifications as nominal-scale. A special case in which $\{\beta_i\}$ in the model have a linear trend is useful when the categories are ordinal. Let $u_1 \leq u_2 \leq \cdots \leq u_c$ denote ordered scores for both the row and column categories. The *ordinal quasi-symmetry model*[8] is

$$\log(\pi_{ij}/\pi_{ji}) = \beta(u_j - u_i). \tag{8.5}$$

This has the form of the usual logistic model, $\text{logit}[\pi(x)] = \alpha + \beta x$, with $\alpha = 0$, $x = u_j - u_i$, and $\pi(x)$ equal to the conditional probability for cell (i, j), given response in cell (i, j) or cell (j, i). To fit the model, we identify (n_{ij}, n_{ji}) as binomial numbers of successes and failures in $n_{ij} + n_{ji}$ trials, and fit a logistic model with intercept forced to equal 0 and with value of the predictor x equal to $u_j - u_i$.

Symmetry and thus marginal homogeneity is the special case $\beta = 0$. An ordinal likelihood-ratio test of marginal homogeneity uses the difference between the deviance values for the symmetry and ordinal quasi-symmetry models, with $df = 1$. The greater the value of $|\beta|$, the greater the difference between π_{ij} and π_{ji} and between the marginal distributions. With scores $\{u_i = i\}$, the probability that the second observation is x categories higher than the first observation equals $\exp(x\beta)$ times the probability that the first observation is x categories higher than the second observation.

For the ordinal quasi-symmetry model with the chosen category scores $\{u_i\}$, the fitted mean for the row variable is $\sum_i u_i \hat{\pi}_{i+}$. This equals the sample row mean, $\sum_i u_i p_{i+}$. A similar equality holds for the column means. When responses in one margin tend to be higher on the ordinal scale than those in the other margin, the fit of this model exhibits this

[7] With the `LORstr="independence"` option in `multgee` for local odds ratios.

[8] Proposed by A. Agresti, *Statist. Prob. Letters* **1**: 313–316 (1983).

same ordering. When $\hat{\beta} > 0$, the mean response is higher for the column variable. When $\hat{\beta} < 0$, the mean response is lower for the column variable.

8.4.4 Example: Recycle or Drive Less Revisited? *

Table 8.6 summarized behaviors on recycling and driving less to help the environment. The symmetry model fits poorly, having deviance 1106.12 with $df = 6$. By comparison, the quasi-symmetry model fits well, having deviance 2.68 with $df = 3$. The simpler ordinal quasi-symmetry model also fits well. For the scores (1, 2, 3, 4), its deviance is 4.41, with $df = 5$, as the next R output shows:

```
-------------------------------------------------------------------------------
> Envir <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Environment.dat",
+                     header=TRUE) # Environment data file at text website
> Envir # nij and nji are opposite cell counts in Table 8.6
  always often sometimes never x nij nji # x=j-i = distance between categories
1      1    -1         0     0 1  43   4
2      1     0        -1     0 2 163   4
3      1     0         0    -1 3 233   0
4      0     1        -1     0 1  99   8
5      0     1         0    -1 2 185   1
6      0     0         1    -1 1 230  18 # 6 pairs of opposite cells
> fit <- glm(nij/(nij+nji) ~ - 1 + x, family=binomial, weights=nij+nji,
+            data=Envir)
> summary(fit) # ordinal quasi symmetry; -1 in model sets intercept = 0
  Estimate  Std. Error  z value  Pr(>|z|)
x   2.3936      0.1508    15.88    <2e-16 # estimate of beta
---
    Null deviance: 1106.1240  on 6  degrees of freedom # null model = symmetry
Residual deviance:    4.4139  on 5  degrees of freedom # ordinal quasi symm.

> fit.QS <- glm(nij/(nij+nji) ~ -1 + always + often + sometimes + never,
+               family=binomial, weights=nij+nji, data=Envir)
> summary(fit.QS) # quasi symmetry
    Null deviance: 1106.1240  on 6  degrees of freedom # symmetry
Residual deviance:    2.6751  on 3  degrees of freedom # quasi symmetry
-------------------------------------------------------------------------------
```

For the ordinal quasi-symmetry model, $\hat{\beta} = 2.394$. From (8.5), the estimated probability that response on driving less is x categories higher than the response on recycling equals $\exp(2.394x)$ times the reverse probability. Responses on recycling tend to be lower on the ordinal scale (i.e., more frequent) than those on driving less. The mean for recycling is 2.1, close to the *often* score, whereas the mean for driving less is 3.5, midway between the *sometimes* and *never* scores.

For these data, the likelihood-ratio statistic for testing marginal homogeneity by comparing the deviances of the symmetry and ordinal quasi-symmetry models is $1106.12 - 4.41 = 1101.71$, with $df = 1$ ($P < 0.0001$). We obtained similar extremely strong evidence of marginal heterogeneity with a cumulative logit marginal model in Section 8.4.2.

8.5 ANALYZING RATER AGREEMENT *

We now present models for data in which each matched pair consists of ratings by two observers on a categorical scale. Such data sets occur frequently in medical applications in which the observers provide diagnoses about some condition. Many categorical scales are quite subjective, and the observers rarely agree perfectly on all cases. This section presents ways to measure strength of agreement and detect patterns of disagreement. *Agreement* is distinct from *association*. Strong agreement requires strong association, but strong association can exist without strong agreement. For an ordinal classification scale, for example, if one observer consistently classifies subjects one level higher than the other observer, the strength of agreement is poor even though the association is strong.

8.5.1 Example: Agreement on Carcinoma Diagnosis

Table 8.7 shows ratings by two pathologists, identified merely as X and Y, who separately classified 118 slides on the presence and extent of carcinoma of the uterine cervix. The rating scale has the ordered categories (1) negative, (2) atypical squamous hyperplasia, (3) carcinoma *in situ*, (4) squamous or invasive carcinoma.

Table 8.7 Diagnoses of carcinoma, with standardized residuals for independence model.

	Pathologist Y				
Pathologist X	1	2	3	4	Total
1	22	2	2	0	26
	(8.5)	(−0.5)	(−5.9)	(−1.8)	
2	5	7	14	0	26
	(−0.5)	(3.2)	(−0.5)	(−1.8)	
3	0	2	36	0	38
	(−4.1)	(−1.2)	(5.5)	(−2.3)	
4	0	1	17	10	28
	(−3.3)	(−1.3)	(0.3)	(5.9)	
Total	27	12	69	10	118

Source: N.S. Holmquist, C.A. McMahon, and O.D. Williams, *Arch. Pathol.* **84**: 334–345 (1967); reprinted with permission by the American Medical Association.

Here π_{ij} is the probability that observer X classifies a slide in category i and observer Y classifies it in category j. In the square contingency table, the main diagonal $\{i = j\}$ represents observer agreement, with π_{ii} being the probability that they both classify a subject in category i. The total probability of agreement is $\sum_i \pi_{ii}$. Perfect agreement occurs when $\sum_i \pi_{ii} = 1$.

8.5.2 Cell Residuals for Independence Model

One way of evaluating agreement compares the cell counts $\{n_{ij}\}$ to the values $\{n_{i+}n_{+j}/n\}$ predicted by the loglinear model of independence (7.1). That model provides a baseline, showing the degree of agreement expected if no association existed between the ratings. Normally it fits poorly if there is even only mild agreement, but its cell standardized residuals (Section 2.4.5) provide information about patterns of agreement and disagreement. Cells

with positive standardized residuals have higher frequencies than expected under independence. Ideally, large positive standardized residuals occur on the main diagonal.

In fact, the independence model fits Table 8.7 poorly (deviance 118.0, $df = 9$). The table reports the standardized residuals in parentheses. The large positive standardized residuals on the main diagonal indicate that agreement for each category is greater than expected by chance, especially for the first category. The off-main-diagonal residuals are primarily negative. Disagreements occurred less than expected under independence, although the evidence of this is weaker for categories closer together. Inspection of cell counts reveals that the most common disagreements refer to observer Y choosing category 3 and observer X instead choosing category 2 or 4.

8.5.3 Quasi-Independence Model

A more useful loglinear model adds a term that describes agreement beyond that expected under independence. This *quasi-independence model* is

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \delta_i I(i = j),$$

where the indicator $I(i = j)$ equals 1 when $i = j$ and equals 0 when $i \neq j$. This model adds to the independence model a parameter δ_1 for cell (1,1) (in row 1 and column 1), a parameter δ_2 for cell (2,2), and so forth. When $\delta_i > 0$, more agreements occur on outcome i than are expected under independence. Because of the addition of this term, the quasi-independence model treats the main diagonal differently from the rest of the table. The *ML* fit in those cells is perfect, with $\hat{\mu}_{ii} = n_{ii}$ for all i. For the remaining cells, the independence model still applies. In other words, conditional on observer disagreement, the rating by X is independent of the rating by Y.

You can fit the quasi-independence model as a loglinear model, as shown in the next R output:

```
-------------------------------------------------------------------------------
> Pathology <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                          Pathologists.dat", header=TRUE)
> Pathology # Data file at text website
   X Y count diag # diag: separate category (1,2,3,4) for each main-diagonal
1  1 1    22    1 #       cell and a single category (0) for all other cells
2  1 2     2    0
3  1 3     2    0
4  1 4     0    0
5  2 1     5    0
6  2 2     7    2
7  2 3    14    0
8  2 4     0    0
9  3 1     0    0
10 3 2     2    0
11 3 3    36    3
12 3 4     0    0
13 4 1     0    0
14 4 2     1    0
```

```
15 4 3    17    0
16 4 4    10    4
> fit <- glm(count ~ factor(X) + factor(Y) + factor(diag),
+            family=poisson, data=Pathology)
> summary(fit) # quasi independence, not showing intercept or main effects
               Estimate  Std. Error  z value  Pr(>|z|)
factor(diag)1    3.8611      0.7297    5.291  1.22e-07
factor(diag)2    0.6042      0.6900    0.876   0.38119
factor(diag)3    1.9025      0.8367    2.274   0.02298
factor(diag)4   20.9877   4988.8789    0.004   0.99664 # ML actually infinite
---
Residual deviance:  13.178  on  5  degrees of freedom
--------------------------------------------------------------------------------
```

The quasi-independence model has deviance 13.18, with $df = 5$. It fits much better than the independence model, but some lack of fit remains.

The quasi-independence model is often inadequate for ordinal scales, which almost always exhibit a positive association between ratings. Conditional on observer disagreement, a tendency usually remains for high (low) ratings by one observer to occur with relatively high (low) ratings by the other observer. The quasi-symmetry model is more complex than the quasi-independence model and often fits much better. It also fits the main diagonal perfectly, but it permits association off the main diagonal. For Table 8.7, it has deviance 1.0, based on $df = 2$.

The symmetry model fits Table 8.7 poorly, with deviance 39.2 ($df = 5$). The deviance difference $39.2 - 1.0 = 38.2$ compared to quasi symmetry, with $df = 3$, provides strong evidence of marginal heterogeneity. The lack of perfect agreement reflects differences in marginal distributions. Table 8.7 reveals these to be substantial in each category but the first. The ordinal quasi-symmetry model also fits poorly, partly because ratings do not tend to be consistently higher by one observer than the other.

8.5.4 Quasi Independence and Odds Ratios Summarizing Agreement

For a pair of subjects, consider the event that each observer classifies one subject in category a and one subject in category b. The odds that the two observers agree rather than disagree on which subject is in category a and which is in category b equal

$$\tau_{ab} = \frac{\pi_{aa}\pi_{bb}}{\pi_{ab}\pi_{ba}}.$$

The further that τ_{ab} is above 1.0, the more likely the observers are to agree on which subject receives each designation.

For the quasi-independence model, the odds summarizing agreement for categories a and b equal

$$\tau_{ab} = \exp(\delta_a + \delta_b).$$

Larger $\{\delta_i\}$ represent stronger agreement. For instance, categories 2 and 3 in Table 8.7 have $\hat{\delta}_2 = 0.60$ and $\hat{\delta}_3 = 1.90$. The estimated odds that one observer's rating is category 2 rather

than 3 are $\hat{\tau}_{23} = \exp(0.60 + 1.90) = 12.3$ times as high when the other observer's rating is 2 than when it is 3. The degree of agreement is quite strong, which also happens for the other pairs of categories.

8.5.5 Kappa Summary Measure of Agreement

An alternative approach describes strength of agreement using a single summary index, rather than a model. The most popular index, *Cohen's kappa*, compares the agreement to that expected if the ratings were independent. The probability of agreement equals $\sum_i \pi_{ii}$. If the observers' ratings were independent, then $\pi_{ii} = \pi_{i+}\pi_{+i}$ and the probability of agreement equals $\sum_i \pi_{i+}\pi_{+i}$. Cohen's kappa is

$$\kappa = \frac{\sum_i \pi_{ii} - \sum_i \pi_{i+}\pi_{+i}}{1 - \sum_i \pi_{i+}\pi_{+i}}.$$

The numerator compares the probability of agreement to that expected under independence. The denominator replaces $\sum_i \pi_{ii}$ by its maximum possible value of 1, corresponding to perfect agreement. Kappa equals 0 when the agreement merely equals that expected under independence, and it equals 1.0 when perfect agreement occurs. The stronger the agreement, for a given pair of marginal distributions, the higher the value of kappa.

For Table 8.7, $\sum_i \hat{\pi}_{ii} = (22 + 7 + 36 + 10)/118 = 0.636$, whereas $\sum_i \hat{\pi}_{i+}\hat{\pi}_{+i} = 0.281$. Sample kappa equals

$$\hat{\kappa} = (0.636 - 0.281)/(1 - 0.281) = 0.49,$$

as shown in the next R output:

```
-------------------------------------------------------------------------------
> library(psych)
> dat <- matrix(Pathology$count, ncol=4, byrow=TRUE) # using Pathology data
> cohen.kappa(dat)                                   # from previous output
                 lower estimate upper
unweighted kappa  0.38     0.49  0.60 # estimate=0.49, 95% CI is (0.38, 0.60)
weighted kappa    0.71     0.78  0.86
-------------------------------------------------------------------------------
```

The difference between the observed agreement and that expected under independence is about 50% of the maximum possible difference.

Kappa treats the variables as nominal scale. When categories are ordered, it treats a disagreement for categories that are close the same as for categories that are far apart. For ordinal scales, a *weighted kappa* extension gives more weight to disagreements for categories that are farther apart.

Controversy surrounds the usefulness of kappa, primarily because its value depends strongly on the marginal distributions. The same diagnostic rating process can yield quite different values of kappa, depending on the proportions of cases of the various types. Models such as quasi-independence and quasi-symmetry more fully describe the structure of agreement and disagreement.

8.6 BRADLEY–TERRY MODEL FOR PAIRED PREFERENCES *

Table 8.8 summarizes results of matches among five professional tennis players between January 2014 and January 2018. For instance, Roger Federer won 6 of the 15 matches that he and Novak Djokovic played. This section presents a model that applies to data of this sort, in which observations consist of pairwise comparisons that result in a preference for one category over another. The fitted model provides a ranking of the players. It also estimates the probabilities of win and of loss for matches between each pair of players.

Table 8.8 Results of 2014–2018 matches for men tennis players.

	Loser				
Winner	Djokovic	Federer	Murray	Nadal	Wawrinka
Djokovic	–	9	14	9	4
Federer	6	–	5	5	7
Murray	3	0	–	2	2
Nadal	2	1	4	–	4
Wawrinka	3	2	2	3	–

Source: Based on information at `www.atpworldtour.com`.

8.6.1 The Bradley–Terry Model and Quasi-Symmetry

The Bradley–Terry model is a logistic model for paired preference data. For Table 8.8, let Π_{ij} denote the probability that player i is the victor when i and j play. The probability that player j wins is $\Pi_{ji} = 1 - \Pi_{ij}$ (ties cannot occur). The model has player parameters $\{\beta_i\}$ such that

$$\text{logit}(\Pi_{ij}) = \log(\Pi_{ij}/\Pi_{ji}) = \beta_i - \beta_j.$$

The probability that player i wins equals $\frac{1}{2}$ when $\beta_i = \beta_j$ and exceeds $\frac{1}{2}$ when $\beta_i > \beta_j$. One parameter is redundant and the model has no intercept.

This logistic model is equivalent to the quasi-symmetry model (8.4). To fit it,[9] we treat each separate pair of cell counts (n_{ij}, n_{ji}) as an independent binomial variate, as Section 8.3.2 described. For instance, from Federer's perspective, the (Federer, Djokovic) results correspond to 6 successes and 9 failures in 15 trials. From the model fit, the estimate of Π_{ij} is

$$\hat{\Pi}_{ij} = \exp(\hat{\beta}_i - \hat{\beta}_j)/[1 + \exp(\hat{\beta}_i - \hat{\beta}_j)].$$

8.6.2 Example: Ranking Men Tennis Players

For Table 8.8, the following `R` output sets $\beta_5 = 0$ for Wawrinka. The parameter estimates indicate that Djokovic and Federer ranked considerably higher than the other players. The model fits adequately (deviance = 4.40, $df = 6$).

[9] In `R`, the `BradleyTerry2` package facilitates fitting the model and its generalizations.

```
------------------------------------------------------------------------------
> Tennis <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Tennis.dat",
+                      header=TRUE)
> Tennis # data file at text website
   Djokovic Federer Murray Nadal Wawrinka nij nji
1         1      -1      0     0        0   9   6 # Djok won 9 lost 6 vs Fed
2         1       0     -1     0        0  14   3
3         1       0      0    -1        0   9   2
4         1       0      0     0       -1   4   3
5         0       1     -1     0        0   5   0 # Federer always beat Murray
6         0       1      0    -1        0   5   1
7         0       1      0     0       -1   7   2
8         0       0      1    -1        0   2   4
9         0       0      1     0       -1   2   2
10        0       0      0     1       -1   4   3
> fit <- glm(nij/(nij+nji) ~ -1 + Djokovic + Federer + Murray + Nadal
+            + Wawrinka, family=binomial, weights=nij+nji, data=Tennis)
> summary(fit)
           Estimate  Std. Error  z value  Pr(>|z|)
Djokovic    1.1761      0.4995    2.354    0.0185
Federer     1.1358      0.5109    2.223    0.0262
Murray     -0.5685      0.5683   -1.000    0.3172
Nadal      -0.0618      0.5149   -0.120    0.9044
Wawrinka        NA          NA       NA        NA
---
    Null deviance: 25.8960  on 10  degrees of freedom
Residual deviance:  4.3958  on  6  degrees of freedom
------------------------------------------------------------------------------
```

The model fit yields estimated probabilities of victory. To illustrate, when Federer plays Djokovic, the estimated probability of a Federer win is

$$\hat{\Pi}_{21} = \frac{\exp(\hat{\beta}_2 - \hat{\beta}_1)}{1 + \exp(\hat{\beta}_2 - \hat{\beta}_1)} = \frac{\exp(1.136 - 1.176)}{1 + \exp(1.136 - 1.176)} = 0.49.$$

For such small data sets, the model smoothing provides estimates that are more sensible than the sample proportions. For instance, Federer beat Murray in all 5 of their matches, but the model estimates the probability of a Federer victory to be 0.85 rather than 1.00.

To check whether the difference between two players is statistically significant, we compare $(\hat{\beta}_i - \hat{\beta}_j)$ to its *SE*. (A simple way to get this in software is to let player j be the baseline with $\hat{\beta}_j = 0$ and then find the *SE* of $\hat{\beta}_i$.) For instance, for comparing Federer and Djokovic, $\hat{\beta}_2 - \hat{\beta}_1 = -0.040$ has $SE = 0.417$, indicating an insignificant difference. A 95% confidence interval for $\beta_2 - \beta_1$ is $-0.040 \pm 1.96(0.417)$, or $(-0.858, 0.778)$. This translates to (0.30, 0.69) for the probability Π_{25} of a Federer win (e.g., $\exp(0.778)/[1 + \exp(0.778)] = 0.69$).

The assumption of independent, identical trials that leads to the binomial distribution and the usual fit of the logistic model is overly simplistic for this application. For instance, the probability Π_{ij} that player i beats player j may vary according to whether the court is clay, grass, or hard, and it would vary over time.

EXERCISES

8.1 Apply the McNemar test to Table 8.3 on smoking and birthweight. Interpret.

8.2 A recent General Social Survey asked subjects whether they believed in heaven and whether they believed in hell. Table 8.9 shows the results.

a. Test the hypothesis that the population proportions answering *yes* were identical for heaven and hell.

b. Find a 90% confidence interval for the difference between the population proportions. Interpret.

c. Estimate and interpret the odds ratio for a logistic model for the probability of a *yes* response as a function of the item (heaven or hell), using the (**i**) marginal model (8.2), (**ii**) subject-specific model (8.3). Why are the estimates different?

Table 8.9 Data from General Social Survey for Exercise 8.2.

Believe in Heaven	Believe in Hell	
	Yes	No
Yes	833	125
No	2	160

8.3 Explain the difference between $\hat{\beta}$ for marginal model (8.2) and $\hat{\beta}$ for subject-specific model (8.3). Illustrate with Table 8.1 on helping the environment.

8.4 Table 8.10 shows results from a matched case-control study. A study of acute myocardial infarction (MI) among Navajo Indians matched 144 victims of MI according to age and gender with 144 individuals free of heart disease. Subjects were then asked whether they had ever been diagnosed as having diabetes. For the logistic model, the odds that a subject with diabetes is an MI case equal $\exp(\beta)$ times the odds that a subject without diabetes is an MI case. Estimate $\exp(\beta)$ and explain why the estimate is valid even though the study observed diabetes, conditional on MI status.

Table 8.10 Previous diagnoses of diabetes for myocardial infarction (MI) case-control pairs.

	MI Cases	
MI Controls	Diabetes	No Diabetes
Diabetes	9	16
No Diabetes	37	82

Source: Coulehan *et al.*, *Amer. J. Public Health* **76**: 412–414 (1986), reprinted with permission by the American Public Health Association.

8.5 Section 8.1.1 presented the large-sample z and chi-squared McNemar test for comparing dependent proportions. The exact P-value uses the binomial distribution. For Table 8.1, consider H_a: $P(Y_1 = 1) > P(Y_2 = 1)$, or equivalently, H_a: $\pi_{12} > \pi_{21}$.

a. Explain why the exact P-value is the binomial probability of at least 132 successes out of 239 trials when the parameter is 0.50. Use software to show that this P-value $= 0.060$.

b. For these data, how is the mid P-value (Section 1.4.3) defined in terms of binomial probabilities? (This P-value $= 0.053$.)

8.6 Explain the following analogy: The McNemar test is to binary response variables as the paired-difference t test is to normally distributed response variables.

8.7 Table 8.11, from the 2016 General Social Survey, reports religious affiliation now and at age 16 for US residents, with categories (1) Protestant, (2) Catholic, (3) Jewish, (4) None or Other.

a. Fit the symmetry model. Use its standardized residuals to analyze transition patterns between pairs of religions.

b. Fit the quasi-symmetry model. Test its fit and test marginal homogeneity by comparing its fit to symmetry. Interpret.

Table 8.11 Data from General Social Survey for Exercise 8.7.

Affiliation at Age 16	Religious Affiliation Now			
	1	2	3	4
1	1136	49	3	288
2	129	590	2	186
3	1	0	42	12
4	81	9	4	157

8.8 Table 8.12, from the 2016 General Social Survey, reports region of residence now and at age 16 for residents of the US.

a. Test marginal homogeneity by comparing fits of two models.

b. Fit the independence model and the quasi-independence (QI) model. Explain why the QI model gives a dramatic improvement in fit. (*Hint*: For the independence model, the standardized residuals are about 40 for the cells on the main diagonal; what happens with these cells for the QI model?)

Table 8.12 Data from General Social Survey for Exercise 8.8.

Residence at Age 16	Residence Now			
	Northeast	Midwest	South	West
Northeast	394	17	81	38
Midwest	8	596	74	59
South	29	32	769	35
West	10	24	35	417

8.9 Table 8.13 is from a General Social Survey. Subjects were asked their opinion about a man and a woman having sex relations before marriage and a married person having sexual relations with someone other than the marriage partner. The response categories are 1 = always wrong, 2 = almost always wrong, 3 = wrong only sometimes, 4 = not wrong at all. The ordinal quasi-symmetry model with scores (1, 2, 3, 4) has deviance 2.1 ($df = 5$) and $\hat{\beta} = -2.86$. Show how to compare to the symmetry model (which has deviance 402.2) to test marginal homogeneity. From $\hat{\beta}$, explain

Table 8.13 Data from General Social Survey for Exercise 8.9.

Premarital Sex	Extramarital Sex			
	1	2	3	4
1	144	2	0	0
2	33	4	2	0
3	84	14	6	1
4	126	29	25	5

why responses on extramarital sex tend to be lower on the ordinal scale than those on premarital sex. (The mean scores are 1.28 for extramarital sex and 2.69 for premarital sex.)

8.10 Table 8.14 is from a General Social Survey. Subjects were asked whether danger to the environment was caused by car pollution and/or by a rise in the world's temperature caused by the *greenhouse effect*. The response categories are 1 = extremely dangerous, 2 = very dangerous, 3 = somewhat dangerous, 4 = not or not very dangerous. Analyze these data by fitting a model, interpreting parameter estimates, and conducting inference. Prepare a short report summarizing your analyses, with edited sofware output as an appendix.

Table 8.14 Data for Exercise 8.10.

Car Pollution	Greenhouse Effect			
	1	2	3	4
1	95	72	32	8
2	66	129	116	13
3	31	101	233	82
4	5	4	24	26

8.11 For Table 8.5 on choice of coffee brand on two occasions, show that the quasi-independence model fits dramatically better than the ordinary independence model. Interpret.

8.12 Table 8.15 displays diagnoses of multiple sclerosis for two neurologists. The categories are (1) Certain multiple sclerosis, (2) Probable multiple sclerosis, (3) Possible

Table 8.15 Data on Multiple Sclerosis Ratings for Exercise 8.12.

Neurologist A	Neurologist B			
	1	2	3	4
1	38	5	0	1
2	33	11	3	0
3	10	14	5	6
4	3	7	3	10

Source: Based on data in J.R. Landis and G. Koch, *Biometrics* **33**: 159–174 (1977). Reprinted with permission from the Biometric Society.

multiple sclerosis, (4) Doubtful, unlikely, or definitely not multiple sclerosis. Analyze the agreement.

8.13 On their TV show, Chicago film critics Gene Siskel and Roger Ebert portrayed themselves as adversarial, with quite different opinions. During a two-year period, the counts of (positive, mixed, negative) Ebert reviews were (64, 9, 10) for positive Siskel reviews, (11, 13, 8) for mixed Siskel reviews, and (13, 8, 24) for negative Siskel reviews.[10] Use models and kappa to summarize their agreement.

8.14 A sample of psychology graduate students at the University of Florida made blind, pairwise preference tests of three cola drinks. For 49 comparisons of Coke and Pepsi, Coke was preferred 29 times. For 47 comparisons of Classic Coke and Pepsi, Classic Coke was preferred 19 times. For 50 comparisons of Coke and Classic Coke, Coke was preferred 31 times. Comparisons resulting in ties are not reported.

a. Fit the Bradley–Terry model and establish a ranking of the drinks.

b. Estimate the probability that Coke is preferred to Pepsi, using the model fit, and compare to the sample proportion.

8.15 Table 8.16 summarizes results of tennis matches for several women professional players between January 2014 and June 2017.

a. Fit the Bradley–Terry model. Report the parameter estimates and rank the players.

b. Estimate and construct a 90% confidence interval for the probability that Serena Williams beats Venus Williams. Interpret.

Table 8.16 Women's tennis data for Exercise 8.15.

	Loser				
Winner	Kerber	Halep	Pliskova	S. Williams	V. Williams
Kerber	–	5	4	1	1
Halep	3	–	5	2	4
Pliskova	3	2	–	1	1
S. Williams	3	5	2	–	3
V. Williams	2	2	1	1	–

Source: Based on information at `www.wtatennis.com`.

8.16 When the Bradley–Terry model holds, explain why it is not possible that A could be preferred to B (i.e., $\Pi_{AB} > \frac{1}{2}$) and B could be preferred to C, yet C could be preferred to A.

8.17 True or false: With positively correlated dependent samples, the estimator of the difference of marginal probabilities is more precise than with independent samples of the same size.

[10] *Source:* A. Agresti and L. Winner, *CHANCE* **10**: 10–14 (1997).

CHAPTER 9

MARGINAL MODELING OF CORRELATED, CLUSTERED RESPONSES

Many studies observe the response variable for each subject repeatedly, at several times (such as in longitudinal studies) or under various conditions. This is common in health-related applications, such as when a physician evaluates patients at regular time intervals regarding whether a drug treatment is successful. Repeated observations on a subject are typically positively correlated. Positive correlations also usually occur when the response variable is observed for *matched sets* of subjects. For example, a study of factors that affect childhood obesity might sample families and then observe the children in each family. A matched set consists of children within a particular family. Children from the same family tend to respond more similarly than children from different families.

We will refer to a matched set of observations as a *cluster*. For repeated measurement of subjects, the set of observations for a particular subject forms a cluster. Statistical analyses should take the correlation into account. In particular, analyses that ignore the correlation can have badly biased standard error estimators. The next two chapters generalize to matched sets the methods of Chapter 8 for matched pairs, while also including explanatory variables.

We first amplify the distinction between *marginal models* and *subject-specific models*. This chapter focuses on marginal models, fitted by solving *generalized estimating equations* (*GEE*). This is a multivariate method that, for discrete data, is computationally much simpler than ML and more readily available with software. We illustrate for binary responses and then for multicategory responses. We also present a *transitional* approach that models observations in a longitudinal study using explanatory variables that include previous

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

response outcomes. Finally, we discuss the impact of clusters missing some observations, which is a common problem with repeated measurement data.

9.1 MARGINAL MODELS VERSUS SUBJECT-SPECIFIC MODELS

Denote the response-variable observations in a cluster by $(y_1, y_2, \ldots, y_T)$, where T denotes the cluster size.[1] As with independent observations, models for correlated, clustered categorical observations focus on how the probability of a particular outcome depends on explanatory variables.

9.1.1 Marginal Models for a Clustered Binary Response

For matched pairs on a binary response, such as we observed in Table 8.1 on environmental opinions, the responses (y_1, y_2) have marginal probabilities for the contingency table that cross-classifies y_1 and y_2. For T repeated binary responses, $\{P(Y_t = 1), t = 1, \ldots, T\}$ are marginal probabilities of a T-dimensional contingency table that cross-classifies them. Marginal models describe how the logits of the marginal probabilities depend on explanatory variables.

For matched pairs, equation (8.2) showed how to construct a logistic model for comparing marginal distributions. An extension of that matched-pairs model allows $T > 2$ observations in each cluster as well as explanatory variables.

9.1.2 Example: Repeated Responses on Similar Survey Questions

In a recent General Social Survey, the subjects indicated whether they supported legalized abortion in each of three situations. Table 9.1 cross-classifies subjects by responses in the three situations and by their gender. A cluster is a set of the three observations for a particular subject. Each gender has a $2\times2\times2$ table for the $T = 3$ responses.

Table 9.1 Support (1 = yes, 0 = no) for legalized abortion in three situations,[a] by gender.

	Sequence of Responses in Three Situations							
Gender	(1,1,1)	(1,1,0)	(0,1,1)	(0,1,0)	(1,0,1)	(1,0,0)	(0,0,1)	(0,0,0)
Male	342	26	6	21	11	32	19	356
Female	440	25	14	18	14	47	22	457

[a] Situations are (1) if the family has a very low income and cannot afford any more children, (2) when the woman is not married and does not want to marry the man, and (3) when the woman wants it for any reason.

For a randomly selected subject, let $P(Y_t = 1)$ denote the probability of supporting legalization in situation t, for $t = 1, 2, 3$. Let $x = 1$ for females and 0 for males. A marginal main-effects model for how Y_t depends on the situation t and gender x is

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta_t + \gamma x.$$

[1] In practice, the number of observations often varies by cluster, but it is simpler to use notation that ignores that.

For identifiability, software imposes a constraint for the situation factor, such as $\beta_1 = 0$. We show R code for fitting this model in Section 9.2.3.

9.1.3 Subject-Specific Models for a Repeated Response

For matched pairs, equation (8.3) presented a different type of model that has probabilities at the subject level. That *subject-specific* model permits heterogeneity among subjects, even at fixed levels of the explanatory variables.

Let y_{it} denote the response outcome for person i in situation t. A subject-specific analog of the marginal model just presented is

$$\text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta_t + \gamma x_i.$$

Each person has their own intercept (α_i), reflecting variability in the probability among people for a particular situation and gender. Calling the model *subject-specific* reflects that the $\{\beta_t\}$ effects are defined conditional on the subject. For example, with $\beta_1 = 0$, for each person, the odds of favoring legalized abortion in situation 2 are $\exp(\beta_2)$ times the odds of favoring it in situation 1. By contrast, the effects in the marginal models specified in the previous subsection are *population-averaged*, because they refer to the entire population rather than to individual subjects. For example, with $\beta_1 = 0$ in the marginal model, in the overall population the odds of favoring legalized abortion in situation 2 are $\exp(\beta_2)$ times the odds of favoring it in situation 1.

The remainder of this chapter focuses only on marginal models. The following chapter presents subject-specific models and also discusses issues relating to the choice of model.

9.2 MARGINAL MODELING: THE GENERALIZED ESTIMATING EQUATIONS (GEE) APPROACH

ML fitting of marginal logit models can be difficult. We will not explore the technical reasons here, but basically it is because the models refer to *marginal* probabilities whereas the likelihood function that is maximized to obtain ML estimates refers to the *joint* distribution of the clustered responses. This section uses a simple alternative to ML that is widely available in software.

9.2.1 Quasi-Likelihood Methods

A GLM specifies a probability distribution for Y and provides a formula for how its mean $E(Y) = \mu$ depends on the explanatory variables by using a link function to connect μ to a linear predictor. The choice of distribution for Y determines the relationship between μ and var(Y). For binary data with success probability π, for example, Y has $E(Y) = \pi$ and var$(Y) = \pi(1 - \pi)$, which is $\mu(1 - \mu)$. For count data with the Poisson distribution, var$(Y) = \mu$.

For a particular link function and linear predictor in a GLM, the choice of a probability distribution for Y determines the likelihood function and thus the ML estimates. As mentioned in Section 3.5, GLM fitting iteratively solves *likelihood equations* to find the ML

estimates. *Quasi-likelihood* methods are a generalization of ordinary likelihood methods that take into account:

- The GLM likelihood equations use the assumed distribution for Y only in terms of how var(Y) depends on μ.
- A generalized set of "quasi-likelihood" estimating equations need not assume a particular distribution for Y but merely how var(Y) depends on μ. The solutions of these estimating equations are the quasi-likelihood estimators.

The quasi-likelihood approach permits departures from the assumptions required by standard distributions. The methods can deal with overdispersion caused by correlated observations or unobserved explanatory variables. For example, one quasi-likelihood method takes the usual formula for how var(Y) depends on μ for a particular distribution, but multiplies it by a constant that is itself estimated using the data. For clustered binary data, the observations within a cluster are likely to be correlated. Therefore, with T observations in a cluster, the variance of the number of successes in the cluster may differ from the variance formula $T\pi(1-\pi)$ for a binomial distribution, which assumes independent trials. The quasi-likelihood approach permits the variance to be some multiple ϕ of the usual variance, that is, $\phi T\pi(1-\pi)$, and estimates ϕ based on the variability in the sample data. Overdispersion occurs when $\phi > 1$.

9.2.2 Generalized Estimating Equation Methodology: Basic Ideas

A computationally simple alternative to ML for clustered categorical data is a multivariate generalization of quasi-likelihood. Rather than assuming a particular multivariate probability distribution for $(Y_1, \ldots, Y_T)$, this method only links each marginal mean $E(Y_t)$ to a linear predictor and provides a guess for the variance–covariance structure of $(Y_1, \ldots, Y_T)$. The method uses the empirical sample variability to generate appropriate standard errors. It is called the *GEE method* because the estimates are solutions of *generalized estimating equations* that are multivariate generalizations of the likelihood equations solved to find ML estimates for GLMs.

The GEE method has the following steps:

- Specify a marginal model for $E(Y_t)$ by selecting a link function and forming a linear predictor.
- Make an assumption about how var(Y_t) depends on $E(Y_t)$. This is typically based on a natural ordinary distribution for Y_t for the type of data (e.g., binomial for binary data).
- Make an educated guess for the correlation structure among $\{Y_t\}$. This is called the *working correlation* matrix.
- The GEE estimates are the solutions of generalized estimating equations, which utilize the marginal model formula and the assumed structure for the variances and correlations.
- Using the empirical variation and correlations in the data, adjust the standard errors to get more robust values that are valid even if you have misspecified the variance–covariance structure.

One possible working correlation has an *exchangeable* structure. This treats $\rho = \text{corr}(Y_s, Y_t)$ as identical, but unknown, for all pairs s and t. Another possibility, often used for time series data, has an *autoregressive* structure. This has the form $\text{corr}(Y_s, Y_t) = \rho^{t-s}$. For example, $\text{corr}(Y_1, Y_2) = \rho$, $\text{corr}(Y_1, Y_3) = \rho^2$, $\text{corr}(Y_1, Y_4) = \rho^3$, ..., with observations farther apart in time being more weakly correlated. The *independence* working correlation structure assumes $\text{corr}(Y_s, Y_t) = 0$ for each pair. With this structure, the resulting model parameter estimates are identical to ML from treating all the observations as independent. At the other extreme, the *unstructured* working correlation matrix permits $\text{corr}(Y_s, Y_t)$ to differ for each pair.

In practice, usually little if any *a priori* information is available about the correlation structure. The lack of assumption needed for the unstructured case seems desirable, but this has the disadvantage of several extra parameters to estimate, especially when T is large. Unless you expect dramatic differences among the correlations, we recommend using the exchangeable working correlation structure. Even if your guess about the correlation structure is poor, valid standard errors result from the adjustment the GEE method makes using the sample covariation. That is, the naive standard errors based on the assumed correlation structure are updated using the empirical evidence. The result is *robust* standard errors that are usually more appropriate than naive ones based solely on the assumed correlation structure. Some software refers to the robust standard errors as *sandwich* standard errors, because the covariance matrix on which they are based uses a formula that sandwiches the empirical information between two matrices that relate to the naive covariance matrix.

9.2.3 Example: Opinion about Legalized Abortion Revisited

For Table 9.1 showing opinions about legalized abortion in three situations, Section 9.1.2 introduced the marginal model

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta_t + \gamma x,$$

where Y_t is the response in situation t for a randomly selected subject. We first perform an analysis that ignores the fact that each person had three observations, instead treating the 3 responses on the 1850 subjects as if they came from a random sample of $3(1850) = 5550$ people. The following R code imposes $\beta_3 = 0$ by forming the factor with situation 3 as the first category. Then, the parameters compare situations 1 and 2 to situation 3, which states that a woman should be able to get an abortion for any reason.

```
------------------------------------------------------------------------------
>  Abortion <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                          Abortion.dat", header=TRUE)
> Abortion # subject-specific data file at text website
 person  gender  situation  response
      1       1          1         1
      1       1          2         1
      1       1          3         1
 ...
   1850       0          1         0
   1850       0          2         0
   1850       0          3         0
```

```
> sit <- factor(Abortion$situation, levels=c(3,1,2))
> fit.glm <- glm(response ~ sit + gender, family=binomial, data=Abortion)
> summary(fit.glm) # ML estimates for 5550 independent observations
              Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -0.12541     0.05560   -2.255    0.0241
sit1           0.14935     0.06582    2.269    0.0233
sit2           0.05202     0.06584    0.790    0.4295
gender         0.00358     0.05414    0.066    0.9472
-------------------------------------------------------------------------
```

The estimates and standard errors are the same as we get with the naive results for the GEE method with *independence* working correlation, shown next:

```
-------------------------------------------------------------------------------
> library(gee)
> fit.gee <- gee(response ~ sit + gender, id=person, family=binomial,
+            corstr="independence", data=Abortion) # cluster on "id" variable
> summary(fit.gee)
              Estimate  Naive S.E.  Naive z  Robust S.E.  Robust z
(Intercept)   -0.12541    0.05562   -2.255      0.06758    -1.856
sit1           0.14935    0.06585    2.268      0.02974     5.022
sit2           0.05202    0.06587    0.790      0.02705     1.923
gender         0.00358    0.05416    0.066      0.08784     0.041
Working Correlation
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
-------------------------------------------------------------------------------
```

The robust standard errors from this GEE analysis are quite different from the naive ones. When positive within-cluster correlation occurs, as often happens in practice and is the case here, standard errors for *between-cluster* effects (such as the comparison of females with males) tend to be larger than when the observations are independent. By contrast, standard errors for *within-cluster* effects, such as comparisons of the different situations, tend to be smaller than when the observations are independent.

The large difference between the naive and robust standard errors reflects the very strong correlations between responses for the different situations. In fact, if we instead use the GEE method with an exchangeable correlation structure, we estimate a common correlation of 0.817 between pairs of responses, as shown in the next R output:

```
-------------------------------------------------------------------------
> fit.gee2 <- gee(response ~ sit + gender, id=person, family=binomial,
+                corstr="exchangeable", data=Abortion)
> summary(fit.gee2)
             Estimate  Naive S.E.  Naive z  Robust S.E.  Robust z
(Intercept)  -0.12533    0.06783   -1.848     0.06758     -1.854
sit1          0.14935    0.02814    5.307     0.02974      5.022
```

```
sit2          0.05202    0.02815     1.848     0.02705      1.923
gender        0.00344    0.08791     0.039     0.08784      0.039
Working Correlation
        [,1]     [,2]     [,3]
[1,] 1.00000  0.81733  0.81733
[2,] 0.81733  1.00000  0.81733
[3,] 0.81733  0.81733  1.00000
------------------------------------------------------------------------------
```

With the exchangeable working correlation, the naive and robust standard errors are not much different. This structure seems to capture well the strong sample correlations. In fact, GEE with unstructured working correlation (not shown here) yields similar values (0.825, 0.796, 0.831) for the three correlations.

Based on the robust SEs, the $\{\hat{\beta}_t\}$ indicate greater support of legalization in situation 1 (when the family has a low income and cannot afford any more children) than the other two, but the effects are weak. For instance, the estimated odds of supporting legalized abortion in situation 1 equal $\exp(0.149) = 1.16$ times the estimated odds in situation 3. Since $\hat{\gamma} = 0.003$, for each situation the estimated probability of supporting legalized abortion is similar for females and males.

To allow interaction between gender and the situation factor, a model uses different $\{\beta_t\}$ for men and women. This corresponds to having extra parameters that are the coefficients of cross-products of the gender and the situation indicator variables. Such a model does not fit significantly better.

9.2.4 Limitations of GEE Compared to ML

The GEE method assumes a probability distribution for each *marginal* distribution, but it makes no assumption about the *joint* distribution of $(Y_1, \ldots, Y_T)$ other than to select a working correlation structure. This is helpful. For continuous multivariate responses it is common to assume a multivariate normal distribution, but for discrete data, such as a categorical response or a count response, no multivariate generalization of standard univariate distributions such as the binomial and Poisson provides simple specification of correlation structure. Because the GEE method does not specify the complete multivariate distribution, it does not have a likelihood function. In this sense, the GEE method is a multivariate type of quasi-likelihood method. Therefore, its estimates are not ML estimates.

For clustered data, the GEE method is much simpler computationally than ML and more readily available in software. However, it has limitations. Because it does not have a likelihood function, there is no deviance, and likelihood-ratio methods are not available for checking fit, comparing models, and conducting inference about parameters. Inference instead uses statistics, such as Wald statistics, based on the approximate normality of the sampling distributions of the estimators together with their robustly estimated standard errors. Such inference is reliable mainly for large samples. Otherwise, the empirically based standard errors tend to underestimate the true ones.

In the opinions about legalized abortion example, suppose we want to test the situation effect, adjusting for gender. This is a generalization of the hypothesis of marginal

homogeneity considered in Chapter 8, where we now have $T = 3$ and an explanatory variable. With GEE methods, we do this with a Wald test, as shown next using R:

```
-------------------------------------------------------------------------
> library(geepack) # geepack library enables Wald tests comparing models
> fit <- geeglm(response ~ gender + factor(situation), id=person,
+               family=binomial, corstr="exchangeable", data=Abortion)
> anova(fit)
Terms added sequentially (first to last) # gender first alone in model
                   Df       X2  P(>|Chi|) # then situation is added
gender              1   0.0017       0.97
factor(situation)   2  26.0171    2.2e-06
-------------------------------------------------------------------------
```

The Wald statistic of 26.02 with $df = 2$ gives strong evidence against the hypothesis of no situation effect, adjusted for gender. With ML for ordinary GLMs, we would do this with a likelihood-ratio test, comparing deviances.

9.3 MARGINAL MODELING FOR CLUSTERED MULTINOMIAL RESPONSES

The GEE method extends to include model-fitting for clustered multicategory response variables. Models for marginal distributions should recognize the measurement scale of the response. With nominal-scale responses, baseline-category logit models describe the odds of each outcome relative to a baseline. We saw an example that applied GEE methods for such models for matched pairs in Section 8.3.2. For ordinal-scale responses, cumulative logit models describe odds for the cumulative probabilities. Section 8.4.2 showed an example for matched pairs and this section shows an example that includes explanatory variables. As in the binary case, the GEE method uses the empirical covariation among the clustered responses to construct robust standard errors.

9.3.1 Example: Insomnia Study

For a sample of patients with insomnia problems, Table 9.2 shows results of a randomized, double-blind clinical trial comparing an active hypnotic drug with a placebo. The response variable is an ordered categorical grouping of the patient's reported time, in minutes, to fall asleep after going to bed. Patients responded before and following a two-week treatment period. The two treatments, active drug and placebo, form a binary explanatory variable. The study randomly allocated the subjects to the treatment groups. Here, each subject forms a cluster, with the observations in a cluster being the ordinal response at the two occasions of observation.

Table 9.3 displays sample marginal distributions for the four treatment $\times$ occasion combinations. From the initial to follow-up occasion, time to falling asleep seems to shift downwards for both treatments. The degree of shift seems greater for the active drug, indicating possible interaction. Let t denote the occasion (0 = initial, 1 = follow-up) and let x denote the treatment (0 = placebo, 1 = active drug). The cumulative logit marginal model

$$\text{logit}[P(Y_t \leq j)] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3 (t \times x)$$

Table 9.2 Time to falling asleep, by treatment and occasion.

		Time to Falling Asleep (minutes)			
	Initial	Follow-up Response			
Treatment	Response	<20	20–30	30–60	>60
Active	<20	7	4	1	0
	20–30	11	5	2	2
	30–60	13	23	3	1
	>60	9	17	13	8
Placebo	<20	7	4	2	1
	20–30	14	5	1	0
	30–60	6	9	18	2
	>60	4	11	14	22

Source: From S.F. Francom, C. Chuang-Stein, and J.R. Landis, *Statist. Med.* **8**: 571–582 (1989). Reprinted with permission from John Wiley & Sons, Ltd.

Table 9.3 Sample marginal distributions of Table 9.2.

		Response			
Treatment	Occasion	<20	20–30	30–60	>60
Active	Initial	0.101	0.168	0.336	0.395
	Follow-up	0.336	0.412	0.160	0.092
Placebo	Initial	0.117	0.167	0.292	0.425
	Follow-up	0.258	0.242	0.292	0.208

permits interaction between occasion and treatment. Like the cumulative logit models of Section 6.2, it makes the *proportional odds* assumption by which the effects are identical for each response cutpoint.

For independence working correlations, the GEE estimates and robust SE values (in parentheses) are

$$\hat{\beta}_1 = 1.038\ (0.168),\ \hat{\beta}_2 = 0.034\ (0.238),\ \hat{\beta}_3 = 0.708\ (0.244).$$

The following `R` output from which these are taken uses the `multgee` package, which expresses pairwise associations in terms of local odds ratios (Section 7.5.1) for pairs of adjacent rows and adjacent columns. Options include *independence* (odds ratios = 1), for which the GEE estimates are identical to ML estimates based on treating observations in a cluster as independent, and *uniform*, which uses a common local odds ratio value for all pairs of observations. The output labels the robust standard errors as *san.se*, where *san* is short for *sandwich* (see Section 9.2.2).

```
------------------------------------------------------------------------------
> Insomnia <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Insomnia.dat",
+                        header=TRUE)
> Insomnia # Insomnia subject-specific data file at text website
```

```
    case     treat  occasion   response
1      1         1         0          1
2      1         1         1          1
...
477  239         0         0          4
478  239         0         1          4
> library(multgee)
> fit <- ordLORgee(response ~ occasion + treat + occasion:treat, id=case,
+        LORstr = "independence", data=Insomnia) # cluster on "id" variable
> summary(fit)
Link : Cumulative logit
Local Odds Ratios Structure: independence
               Estimate   san.se    san.z  Pr(>|san.z|) # san = sandwich
occasion        1.03808  0.16759   6.1943       < 2e-16 # not showing
treat           0.03361  0.23844   0.1410       0.88790 # 3 intercepts
occasion:treat  0.70776  0.24352   2.9064       0.00366
------------------------------------------------------------------------------
```

The test statistic $z = \hat{\beta}_3/SE = 0.708/0.244 = 2.91$ provides strong evidence of interaction (P-value $= 0.004$). At the initial observation, the estimated odds that time to falling asleep for the active drug is below any fixed level equal $\exp(0.034) = 1.03$ times the estimated odds for the placebo treatment. In other words, initially the two groups had similar distributions, as expected by the randomization of subjects to treatments. At the follow-up observation, the effect is $\exp(0.034 + 0.708) = 2.10$. Those taking the active drug tended to fall asleep more quickly than those taking placebo.

For a simpler interpretation, it can be helpful to assign scores to the ordered categories and report the sample marginal means and their differences. With response scores (10, 25, 45, 75) for time to fall asleep, the initial means were 50.0 for the active drug and 50.3 for the placebo. The difference in means between the initial and follow-up responses was 22.2 for the active drug and 13.0 for the placebo.

If we had naively treated repeated responses as independent for the entire analysis, we would have obtained the same estimates as in the GEE analysis but the SE values for within-subject time effects would have been misleadingly large. For example, the interaction effect estimate of 0.708 would have had an SE of 0.334 rather than 0.244.

9.3.2 Alternative GEE Specification of Working Association

For categorical data, specifying working correlations for the clustered responses is a somewhat awkward aspect of GEE. For binary responses, unlike continuous responses, correlations cannot take value over the entire $[-1, +1]$ range. The actual range depends on the marginal probabilities. For multinomial responses, it is even more awkward to use correlations.

The odds ratio is a more suitable measure of the association for categorical response variables. In fact, the `multgee` package used in this section for multinomial responses expresses pairwise associations in terms of local odds ratios.[2] An alternative version of GEE for binary

[2] The package by A. Touloumis is based on an article by him with A. Agresti and M. Kateri, *Biometrics* **69**: 633–640 (2013).

data also can specify working associations using the odds ratio. For example, the exchangeable structure states that the odds ratio is the same for each pair of observations. Some software[3] can provide this version of GEE. Substantive results about model parameters are usually similar to those based on working correlations.

9.4 TRANSITIONAL MODELING, GIVEN THE PAST

When observations occur over time, some studies focus on the dependence of Y_t on the previously observed responses $\{y_1, y_2, \ldots, y_{t-1}\}$ as well as the ordinary explanatory variables. Models that include past observations as explanatory variables are called *transitional models*. They are often relevant when a primary goal is to predict a future response based on responses observed so far, such as in economic projections.

A (first-order) *Markov model* is a transitional model for which, for all t, the conditional distribution of Y_t, given $y_1, \ldots, y_{t-1}$, is identical to the conditional distribution of Y_t given y_{t-1} alone. That is, given y_{t-1}, Y_t is conditionally independent of $Y_1, \ldots, Y_{t-2}$. Knowing the most recent observation, information about previous observations before it does not help with predicting the next observation. A Markov model is adequate for modeling Y_t if the model with y_{t-1} as the only past observation used as an explanatory variable fits as well, for practical purposes, as a model with $\{y_1, y_2, \ldots, y_{t-1}\}$ as explanatory variables.

9.4.1 Transitional Models with Explanatory Variables

With binary response variable y and p ordinary explanatory variables, a *Markov logistic regression model* has the form, for each t,

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta y_{t-1} + \beta_1 x_{1t} + \cdots + \beta_p x_{pt},$$

where x_{jt} is the value of explanatory variable x_j at time t. Given the previous response and the explanatory variables, the model treats repeated observations by a subject as independent. Thus, you can fit the model with ordinary GLM software, treating each observation as separate.

Effects $\{\beta_j\}$ in transitional models differ from effects in marginal models, both in magnitude and in their interpretation. The effect of x_j on Y_t is conditional on y_{t-1} in a transitional model, but it ignores y_{t-1} in a marginal model. Effects in transitional models are often considerably weaker than effects in marginal models, because conditioning on y_{t-1} attenuates the effect.

An explanatory variable in a transitional model may take a different value for each t or it may be constant. For example, in a longitudinal medical study, a subject's blood pressure could change over time but race would not. Higher-order Markov models also include in the predictor set y_{t-2} and possibly other previous observations.

9.4.2 Example: Respiratory Illness and Maternal Smoking

Table 9.4 is from the Harvard study of air pollution and health. At ages 7–10, children were evaluated annually on whether they had a respiratory illness. Explanatory variables are the

[3] Such as SAS (PROC GENMOD), with option *logor* = *exch* instead of *type* = *exch*.

Table 9.4 Child's respiratory illness by age and maternal smoking.

Child's Respiratory Illness			No Maternal Smoking		Maternal Smoking	
Age 7	Age 8	Age 9	Age 10		Age 10	
			No	Yes	No	Yes
No	No	No	237	10	118	6
		Yes	15	4	8	2
	Yes	No	16	2	11	1
		Yes	7	3	6	4
Yes	No	No	24	3	7	3
		Yes	3	2	3	1
	Yes	No	6	2	4	2
		Yes	5	11	4	7

Source: Thanks to the late Dr. James Ware for these data.

age of the child t ($t = 7, 8, 9, 10$) and maternal smoking at the start of the study ($s = 1$ for smoking regularly, $s = 0$ otherwise).

For the response on respiratory illness at age t, in the Markov logistic model

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta y_{t-1} + \beta_1 s + \beta_2 t, \quad t = 8, 9, 10,$$

each subject contributes three observations to the model fitting. The data set consists of 12 binomials, for the $2\times3\times2$ combinations of (s, t, y_{t-1}). For instance, for the combination (0, 8, 0), from Table 9.4 we see that $y_8 = 0$ for $237 + 10 + 15 + 4 = 266$ subjects and $y_8 = 1$ for $16 + 2 + 7 + 3 = 28$ subjects.

The ML fit of this Markov logistic model is

$$\text{logit}[\hat{P}(Y_t = 1)] = -0.293 + 2.211 y_{t-1} + 0.296 s - 0.243 t.$$

Here is how we can obtain this with ordinary logistic fitting in R:

```
------------------------------------------------------------------------------
> Respiratory <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                           Respiratory.dat", header=TRUE)
> Respiratory # Respiratory data file at text website
    no yes previous s  t
1  283  17        0 0 10
2   30  20        1 0 10
3  140  12        0 1 10
4   21  14        1 1 10
5  274  24        0 0  9
6   26  26        1 0  9
7  134  14        0 1  9
8   18  21        1 1  9
9  266  28        0 0  8
```

```
10  32  24        1 0  8
11 134  22        0 1  8
12  14  17        1 1  8
> fit <- glm(yes/(no+yes) ~ previous + s + t, family=binomial,
+            weights=no+yes, data=Respiratory)
> summary(fit)
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -0.29256     0.84603   -0.346    0.7295
previous      2.21107     0.15819   13.977    <2e-16
s             0.29596     0.15634    1.893    0.0583
t            -0.24281     0.09466   -2.565    0.0103
---
    Null deviance: 207.2212  on 11  degrees of freedom
Residual deviance:   3.1186  on  8  degrees of freedom
> library(car)
> Anova(fit) # likelihood-ratio tests of effects
          LR Chisq  Df  Pr(>Chisq)
previous   192.331   1     < 2e-16
s            3.546   1     0.05969 .
t            6.649   1     0.00992
--------------------------------------------------------------------------
```

Not surprisingly, the previous response y_{t-1} has a strong effect. The multiplicative impact on the odds is $e^{2.211} = 9.1$. Given y_{t-1} and the child's age, there is slight evidence of a positive effect of maternal smoking: the likelihood-ratio statistic for H_0: $\beta_1 = 0$ is 3.55 ($df = 1$, $P = 0.06$). The maternal smoking effect weakens further if we add y_{t-2} to the model (Exercise 9.12).

9.4.3 Group Comparisons Treating Initial Response as a Covariate

A transitional model can be especially useful for matched-pairs data. The marginal models that are the main focus of this chapter evaluate how the marginal distributions of Y_1 and Y_2 depend on explanatory variables. It is often more relevant to treat Y_2 as a univariate response, evaluating effects of explanatory variables while adjusting for the initial response y_1, which is the focus of a transitional model.

We illustrate with the insomnia clinical trial (Section 9.3.1). Let y_1 be the initial time to fall asleep, let Y_2 be the follow-up time, with explanatory variable x defining the two treatment groups (1 = active drug, 0 = placebo). We now treat Y_2 as an ordinal response and y_1 as a quantitative explanatory covariate, using scores (10, 25, 45, 75) for its four categories. In the model

$$\text{logit}[P(Y_2 \le j)] = \alpha_j + \beta_1 x + \beta_2 y_1, \tag{9.1}$$

β_1 compares the follow-up distributions for the two treatments, adjusting for the initial observation. This models the follow-up response (Y_2), conditional on y_1, rather than marginal distributions of (Y_1, Y_2).

From software for ordinary cumulative logit models illustrated in the following R output, the ML treatment effect estimate of $\hat{\beta}_1 = 0.885$ ($SE = 0.246$) provides strong evidence

that follow-up time to fall asleep is lower for the active drug group. Adjusting for the initial response, the estimated odds of falling asleep by a particular time for the active drug are $\exp(0.885) = 2.42$ times those for the placebo group.

```
-------------------------------------------------------------------------------
> Insomnia2 <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Insomnia2.dat",
+                         header=TRUE)
> Insomnia2 # Insomnia2 data file at text website
  treatment initial follow1 follow2 follow3 follow4
1         1      10       7       4       1       0
2         1      25      11       5       2       2
3         1      45      13      23       3       1
4         1      75       9      17      13       8
5         0      10       7       4       2       1
6         0      25      14       5       1       0
7         0      45       6       9      18       2
8         0      75       4      11      14      22
> library(VGAM)
> fit <- vglm(cbind(follow1, follow2, follow3, follow4) ~ treatment + initial,
+             family=cumulative(parallel=TRUE), data=Insomnia2)
> summary(fit)
                Estimate  Std. Error  z value  Pr(>|z|)
(Intercept):1    0.58116     0.31882    1.823  0.068330
(Intercept):2    2.27744     0.35504    6.415  1.41e-10
(Intercept):3    3.75095     0.39980    9.382   < 2e-16
treatment        0.88468     0.24555    3.603  0.000315
initial         -0.04211     0.00580   -7.264  3.75e-13
-------------------------------------------------------------------------------
```

9.5 DEALING WITH MISSING DATA *

Studies with repeated measurement often have cases for which at least one observation in a cluster is missing. In a longitudinal study, some subjects may drop out before the study's end, perhaps because of moving to another city or having some reason they no longer want to participate. Missing data can also be problematic in studies not having repeated measurement. In sample surveys, for example, some subjects may refuse to answer some questions.

A statistical analysis that deletes from the data file all subjects for whom data are missing on at least one variable is called a *complete-case analysis*. This approach can lose a lot of information and result in larger standard errors than if we could use all the available data. Also, the subjects who have some missing data may tend to be systematically different in some way from the other subjects. Because of this, analyzing the observed data alone as if no data are missing can result in biased parameter estimates.

9.5.1 Missing at Random: Impact on ML and GEE Methods

Bias does not occur with complete-case analyses in the rather stringent case in which the data are *missing completely at random* (MCAR). Whether an observation is missing is

statistically independent of that observation's value and the values of the other variables in the entire data file. The cases with missing data are then like a simple random sample of all the cases. A less stringent scenario is that data are *missing at random* (MAR). What caused the data to be missing does not depend on their values. This is true, for example, if whether someone drops out of the study may depend on values observed prior to the drop-out but not on the later unobserved values.

Suppose that a longitudinal study models the response of whether subjects show symptoms of mental depression, with explanatory variables that include a depression severity index measured at the start of the study, treatment type, and the amount of time since treatment began. If the probability that a response is missing is the same for all subjects regardless of values of the explanatory variables, then the data are MCAR. If the probability that the response is missing varies according to time but does not vary according to the response for subjects at the same values of the explanatory variables, then the data are not MCAR but are MAR. They are not MAR if those with a missing response are more likely to be depressed than those not missing it, controlling for the other variables. They are not MAR if whether the response is missing is associated with whether a person is married, and marital status is associated with the response but was not measured in the study.

In practice, we do not know whether MCAR or MAR is satisfied, because we cannot observe the missing data. However, certain evidence can show that they are not satisfied. For example, suppose the proportion of missing responses is much higher for subjects with a severe than with a mild initial depression index. Then, the missing data do not seem to be MCAR, because the missing observations do not resemble a random sample of all the observations.

With clustered data, the clusters can have different numbers of observations. The data file has a separate line for each observation in a cluster, so software does not restrict us to using complete cases. To avoid bias, however, ordinary GEE methods for clustered, correlated data require the MCAR assumption. ML analyses, such as those in earlier chapters and ones we present in the next chapter for random effects models, require the less stringent MAR to avoid bias.

9.5.2 Multiple Imputation: Monte Carlo Prediction of Missing Data

Multiple imputation[4] is a way to predict values for the missing data and use that information in statistical analyses. An imputation is a Monte Carlo method in which the missing values are replaced by simulated versions from their conditional distribution, given the observed data. For example, to impute missing values on a quantitative explanatory variable x_1, you fit a regression model predicting x_1 from $(y, x_2, \ldots, x_p)$ using data available on all of these variables. Then for each subject who is missing the value of x_1, you randomly generate an x_1 value from a normal distribution with mean equal to the predicted value from the prediction equation and with standard deviation equal to the residual estimate for that fitted model. This is a prediction for the missing value that also recognizes how observations vary around their expected values. If x_1 is a binary explanatory variable, we would use logistic regression to obtain a predicted probability for a missing observation, and then generate a 1 or 0 randomly according to a binomial observation with that probability.

[4] The statistician Donald Rubin developed this method in 1987 for nonresponse in surveys.

This imputation is done for every missing observation on every variable, to generate a complete data set for fitting the ordinary model. The *multiple* adjective of multiple imputation means that the imputation process is repeated several times, to provide an indication of how much the resulting model parameter estimates vary from simulation to simulation. Let M denote the number of imputed data sets. The M results are averaged to estimate what we would have found if no data were missing. The resulting estimates have standard errors based on the within-imputation and between-imputation variances, thus incorporating the missing-data uncertainty. Multiple imputation produces more-efficient estimators than analyses that delete observations with missing data. The estimators are not biased when the data are MAR or MCAR. Software can perform the multiple imputation method for you.[5]

When missing data are common, conduct analyses with caution. You should compare results using all available cases to results using only clusters having no missing observations. If you have used multiple imputation, compare to those results also. If results differ substantially, conclusions should be tentative until the reasons for the missingness can be studied. When missingness is not MCAR or MAR, more complex analyses are needed that model the joint distribution of the responses and the binary outcome for each potential observation on whether the observation is missing. Ways to do this and examples implementing multiple imputation are beyond the scope of this book. For details, see Molenberghs and Verbeke (2005) and Molenberghs et al. (2015).

EXERCISES

9.1 Refer to Table 7.1 on high school students' use of alcohol, cigarettes, and marijuana. Consider the data as matched triplets.

a. Construct the marginal distribution for each substance. Specify a marginal model that can compare them and explain how to interpret its parameters. State the hypothesis of marginal homogeneity in terms of model parameters.

b. Specify a corresponding subject-specific model. Explain how parameter interpretation differs from the marginal model.

c. Construct a data file to fit the marginal model. (*Hint*: You could mimic the `Abortion` data file, ignoring gender.) Use GEE methods to fit the model and to test the hypothesis of marginal homogeneity.

9.2 Analyze Table 9.4 using a marginal model with age and maternal smoking as explanatory variables. Report the prediction equation and compare interpretations to the Markov logistic model of Section 9.4.2.

9.3 For Table 7.8, in fitting marginal models to describe main effects of race, gender, and substance type on whether a subject had used that substance, we find evidence of an interaction between gender and substance type. Using GEE with exchangeable working correlation, the estimated probability $\hat{\pi}$ of using a particular substance satisfies

$$\text{logit}(\hat{\pi}) = -0.57 + 1.93s_1 + 0.86s_2 + 0.38r - 0.20g + 0.37g \times s_1 + 0.22g \times s_2,$$

[5] For example, using the `mice` package in `R`, PROC MI in SAS, and the `mi` command in Stata.

where r, g, s_1, s_2 are indicator variables for race (1 = white, 0 = nonwhite), gender (1 = female, 0 = male), and substance type ($s_1 = 1$, $s_2 = 0$ for alcohol; $s_1 = 0$, $s_2 = 1$ for cigarettes; $s_1 = s_2 = 0$ for marijuana). Show that:

a. The group with the highest estimated probability of use of marijuana is white males. What group is it for use of alcohol? cigarettes?

b. Given gender, the estimated odds a white subject used a given substance are 1.46 times the estimated odds for a nonwhite subject.

c. Given race, the estimated odds that a female used (alcohol, cigarettes, marijuana) are (1.19, 1.02, 0.82) times the estimated odds for males.

d. Given race, the estimated odds that a female used (alcohol, cigarettes) are (9.97, 2.94) times the estimated odds she used marijuana.

e. Given race, the estimated odds that a male used (alcohol, cigarettes) are (6.89, 2.36) times the estimated odds he used marijuana.

9.4 Table 9.5 refers to a clinical trial comparing a new drug ($d = 1$) with a standard drug ($d = 0$) for subjects suffering mental depression. Initial severity of depression was recorded as mild ($s = 0$) or severe ($s = 1$). In each group, subjects were randomly assigned to one of the two drugs. Following $t = 1$ week, 2 weeks, and 4 weeks of treatment, each subject's level of mental depression was classified as normal ($y_t = 1$) or abnormal ($y_t = 0$).

a. With scores (0, 1, 2) for t and exchangeable working correlation and permitting a $d \times t$ interaction, we obtain the GEE results shown next. Find and interpret the estimated time effect for each drug.

```
----------------------------------------------------------------------
             Estimate  Naive S.E.  Naive z  Robust S.E. Robust z
(Intercept) -0.02810     0.16255 -0.17286      0.17418 -0.16132
severity    -1.31391     0.14486 -9.07004      0.14596 -9.00167
drug        -0.05927     0.22053 -0.26874      0.22856 -0.25931
time         0.48246     0.11412  4.22786      0.11994  4.02260
drug:time    1.01719     0.18771  5.41910      0.18770  5.41921
----------------------------------------------------------------------
```

b. Explain why at time t, the estimated odds of normal response with the new drug are $\exp(-0.059 + 1.017t)$ times the estimated odds for the standard drug, for each initial diagnosis level. Interpret.

Table 9.5 Depression at three times (N = normal, A = abnormal) by treatment and diagnosis severity.

Diagnosis Severity	Treatment	Response at Three Times							
		NNN	NNA	NAN	NAA	ANN	ANA	AAN	AAA
Mild	Standard	16	13	9	3	14	4	15	6
Mild	New drug	31	0	6	0	22	2	9	0
Severe	Standard	2	2	8	9	9	15	27	28
Severe	New drug	7	2	5	2	31	5	32	6

Source: G. Koch et al. *Biometrics* **33**: 133–158 (1977). Ungrouped data are in `Depression` data file at text website.

9.5 Refer to the previous exercise. Analyze the `Depression` data file at the text website using GEE, assuming exchangeable correlation and with the time scores (1, 2, 4). Interpret model parameter estimates and compare substantive results to those with scores (0, 1, 2).

9.6 Table 9.6 refers to a three-period crossover trial to compare placebo (treatment A), a low-dose analgesic (B), and a high-dose analgesic (C) for relief of primary dysmenorrhea. Subjects were divided randomly into the six possible sequences for administering the treatments. At the end of each period, each subject rated the treatment as giving no relief (0) or some relief (1). Let $y_{i(k)t} = 1$ denote relief for subject i using treatment t ($t = A, B, C$), where i is nested in the treatment sequence k ($k = 1, ..., 6$). Use GEE to find $\{\hat{\beta}_t\}$ for the model

$$\text{logit}[P(Y_{i(k)t} = 1)] = \alpha_k + \beta_t.$$

How would you order the drugs, taking significance into account?

Table 9.6 Crossover data for Exercise 9.6.

Treatment Sequence	Response Pattern for Treatments (A, B, C)							
	000	001	010	011	100	101	110	111
A B C	0	2	2	9	0	0	1	1
A C B	2	0	0	9	1	0	0	4
B A C	0	1	1	8	1	3	0	1
B C A	0	1	1	8	1	0	0	1
C A B	3	0	0	7	0	1	2	1
C B A	1	5	0	4	0	3	1	0

Source: B. Jones and M.G. Kenward, *Statist. Medic.* **6**: 171–181 (1987).

9.7 A Kansas State University survey[6] of 262 pig farmers asked "What are your primary sources of veterinary information?" The farmers were asked to select all relevant categories from (A) Professional Consultant, (B) Veterinarian, (C) State or Local Extension Service, (D) Magazines, and (E) Feed Companies and Reps. The `Farmers` data file at the text website shows results, listing the response for each source (1 = yes, 0 = no) as well as whether the farmer had at least some college education (1 = yes, 0 = no) and size of farm (number of pigs marketed annually, in thousands, with categories 1, 2, 3, 4 for < 1, 1–2, 2–5, > 5). In contingency table form, the data are a $2^5 \times 2 \times 4$ table of (yes, no) counts for each of the five sources cross-classified with the farmers' education and size of farm.

a. Explain why it is not proper to analyze the data by fitting a multinomial model to the counts in the $5 \times 2 \times 4$ contingency table that cross-classifies information source by education and size of farm, treating source as the response variable. (*Hint*: This table contains 453 positive responses of sources from the 262 farmers.)

[6] Thanks to Professor Tom Loughin for showing me these data.

b. For a farmer with education i and size of farm s, let y_{ist} be the response on source t (1 = yes, 0 = no). Using GEE with exchangeable working correlation, estimate parameters and interpret effects in the model lacking an education effect,

$$\text{logit}[P(Y_{ist} = 1)] = \alpha_t + \beta_t s, \quad s = 1, 2, 3, 4.$$

9.8 Table 9.7 is from a longitudinal study of coronary risk factors in school children. A sample of children aged 10–13 were classified by gender and by relative weight (O = obese, N = not obese) at the first observation, two years later, and four years later. Analyze these data. Summarize results in a short report.

Table 9.7 Children classified by gender and relative weight.

	Responses in Three Years							
Gender	NNN	NNO	NON	NOO	ONN	ONO	OON	OOO
Male	119	7	8	3	13	4	11	16
Female	129	8	7	9	6	2	7	14

Source: From R.F. Woolson and W.R. Clarke, *J. Roy. Statist. Soc.* **A 147**: 87–99 (1984). Reproduced with permission from the Royal Statistical Society, London.

9.9 Analyze the 3×3×3×3 table on government spending in Table 7.16 with a marginal cumulative logit model. Interpret effects and show how to test marginal homogeneity.

9.10 Refer to the cereal diet and cholesterol study of Table 6.10 (Exercise 6.16). Analyze these data with ordinal marginal models. Summarize results in a short report, showing edited software output as an appendix.

9.11 For the insomnia study (Table 9.2), the transitional model (9.1) compared treatments while controlling for initial time to fall asleep.

a. Add an interaction term to model (9.1). Summarize how the estimated treatment effect varies according to the initial responses.

b. Now treat the initial response as qualitative. Find and interpret the estimated treatment log odds ratio, fitting models that (i) assume no interaction, (ii) allow interaction. Interpret.

9.12 Analyze Table 9.4 from Section 9.4.2 using a transitional model with *two* previous responses. Given that y_{t-1} is in the model, does y_{t-2} provide additional predictive power? How does the maternal smoking effect compare to the model using only y_{t-1} of the past responses?

9.13 What is wrong with this statement?: "For a first-order Markov logistic regression model, Y_t is independent of Y_{t-2}."

9.14 Give an example of a scenario in which data would likely (**a**) be missing at random but not missing completely at random, (**b**) not be missing at random.

9.15 True or false? For repeated measurement data, under an assumption of independent observations, the ML estimates are valid but their ordinary standard errors are usually too large for within-subject effects and too small for between-subjects effects.

CHAPTER 10

RANDOM EFFECTS: GENERALIZED LINEAR MIXED MODELS

Chapter 9 focused on *marginal modeling* of clustered, correlated categorical responses, such as occur in longitudinal studies. This chapter presents an alternative model type that has a term in the model for each cluster. The *cluster-specific* term, referred to as *subject-specific* in the common application in which each cluster is a person, takes the same value for each observation in a cluster. The model effects have *conditional* interpretations, applying conditionally on the cluster rather than marginally over clusters.

Models usually treat the cluster-specific term as varying randomly among clusters, in which case it is called a *random effect*. *Generalized linear mixed models* extend generalized linear models to include random effects. We show examples of binary and multicategory logistic regression models with random effects and we make comparisons with marginal models. We also present examples of models with multiple random effect terms, such as *multilevel models* that have random effects at different levels of a hierarchy. Random effects are unobserved latent variables, and we also introduce a *latent class model* that has a qualitative latent variable.

10.1 RANDOM EFFECTS MODELING OF CLUSTERED CATEGORICAL DATA

Parameters that describe the effects of a factor in ordinary GLMs are called *fixed effects*. They apply to *all* categories of interest, such as genders or treatments. By contrast, *random effects* apply to a *sample* of clusters from all the possible clusters. For a study with repeated

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

measurement of subjects, for example, a cluster is a set of observations for a particular subject. The model contains a random effect term for each subject and treats those random effects as independently generated from some probability distribution, typically a normal distribution with unknown variance.

10.1.1 The Generalized Linear Mixed Model (GLMM)

Generalized linear models (GLMs) extend ordinary regression by allowing nonnormal responses and a link function of the mean. The *generalized linear mixed model*, denoted by *GLMM*, is a further extension that permits random effects as well as fixed effects in the linear predictor. Denote[1] the random effect for cluster i by u_i. In the most common case, u_i is an intercept term in the model.

Let y_{it} denote observation t in cluster i, such as the response at time t. Let x_{itk} be the value of explanatory variable k for that observation, for $k = 1, \ldots, p$ explanatory variables. Conditional on u_i, a GLMM resembles an ordinary GLM. Let $\mu_{it} = E(Y_{it} \mid u_i)$, the mean of the response variable for a given value of the random effect. With link function g, the most common GLMM has the form

$$g(\mu_{it}) = u_i + \alpha + \beta_1 x_{it1} + \cdots + \beta_p x_{itp}, \;\; i = 1, \ldots, n, \; t = 1, \ldots, T. \quad (10.1)$$

This GLMM, having random effect as part of an intercept term, is called a *random intercept model*. The random effects $\{u_i\}$ are unobserved, treated as independent random variables assumed to have a normal $N(0, \sigma^2)$ distribution. The variance σ^2 is also a parameter, called a *variance component*.

Why not also treat $\{u_i\}$ as fixed effects (i.e., parameters)? The observed clusters are usually a sample of all clusters of interest and treating them as random effects enables us to make inferences to the population of all clusters. In addition, usually a study samples a very large number of clusters and so if treated as fixed effects, the model then contains a very large number of parameters. When we instead treat $\{u_i\}$ as random effects, the model has only a single additional parameter (σ^2), describing the variability among clusters. In some studies, this variability might represent heterogeneity caused by not including certain explanatory variables that are associated with the response variable. The random effect then reflects terms that would be in the fixed effects part of the model if those explanatory variables had been included.

As in ordinary GLMs, ML fitting of GLMMs treats the observations as independent. This independence is assumed conditional on the $\{u_i\}$ as well as the explanatory variable values. In practice, unlike the explanatory variable values, the values of $\{u_i\}$ are unknown. Averaged with respect to the probability distribution of $\{u_i\}$, the model implies nonnegative correlation among observations within a cluster, as discussed in the next subsection. The model-fitting process estimates the fixed effects $(\alpha, \beta_1, \ldots, \beta_p)$ and the variance component $\sigma^2 = \text{var}(u_i)$. It can also provide predictions $\{\hat{u}_i\}$ of the random effects and substitute them and the fixed effect estimates in the linear predictor to estimate the means $\{\mu_{it}\}$.

10.1.2 A Logistic GLMM for Binary Matched Pairs

We illustrate the GLMM expression (10.1) using a simple case — binary matched pairs – which was introduced in Section 8.2.4. Cluster i consists of the observations (y_{i1}, y_{i2}) for

[1] We use a Roman, rather than Greek, letter to emphasize it is a random variable, not a parameter.

matched pair i. Observation t in cluster i has $y_{it} = 1$ (a success) or 0 (a failure), for $t = 1, 2$. The model is

$$\text{logit}[P(Y_{i1} = 1)] = u_i + \alpha + \beta, \quad \text{logit}[P(Y_{i2} = 1)] = u_i + \alpha, \tag{10.2}$$

where $\{u_i\}$ have a $N(0, \sigma^2)$ distribution. Logistic regression models that contain a random effect that is assumed to have a normal distribution are called *logistic-normal models*. This model is the special case of the GLMM (10.1) with $g =$ logit link function, $T = 2$, and an indicator explanatory variable that is 1 for $t = 1$ and 0 for $t = 2$. The fixed effect β represents a log odds ratio at the cluster level. For each cluster, e^β is the odds ratio comparing the response distribution at $t = 1$ with the response distribution at $t = 2$. Section 8.2.4 introduced a similar model with a fixed effect α_i in place of $u_i + \alpha$.

A logistic-normal model with random intercept implies nonnegative correlations among observations within a cluster. This reflects that observations from the same cluster usually tend to be more alike than observations from different clusters. As u_i becomes larger in the positive direction, the probability increases of outcomes $(y_{i1} = 1, y_{i2} = 1)$; as u_i becomes larger in the negative direction, the probability increases of outcomes $(y_{i1} = 0, y_{i2} = 0)$. When a high proportion of cases have these outcomes, the log odds ratio between the pair of repeated responses is positive. Greater association results from greater heterogeneity (i.e., larger σ^2).

10.1.3 Example: Environmental Opinions Revisited

Table 10.1 shows a 2×2 table from the General Social Survey, analyzed previously in Section 8.2. Subjects were asked whether, to help the environment, they were willing to (1) raise taxes, (2) accept a cut in living standards.

Table 10.1 Opinions relating to the environment.

Pay Higher Taxes	Cut Living Standards: Yes	Cut Living Standards: No	Total
Yes	227	132	359
No	107	678	785
Total	334	810	1144

The ML fit of model (10.2), treating $\{u_i\}$ as independent from a $N(0, \sigma^2)$ distribution, yields $\hat{\beta} = 0.210$ $(SE = 0.130)$, with $\hat{\sigma} = 2.85$. For each subject, the estimated odds of a *yes* response on paying higher taxes equal $\exp(0.210) = 1.23$ times the odds of a *yes* response on cutting living standards. Here is R output with these results:

```
------------------------------------------------------------------------------
> Opinions  <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Opinions.dat",
+                         header=TRUE) # Opinions data file at text website
> Opinions
     person question y
1         1        1 1
2         1        0 1
...
2287   1144        1 0
2288   1144        0 0
```

```
> library(lme4) # Fit GLMM by adaptive Gaussian quadrature, with
                # nAGQ quadrature points, as Section 10.1.5 explains
> fit <- glmer(y ~ (1|person) + question, family=binomial, nAGQ=50,
+              data=Opinions) # (1|person) is random intercept for person
> summary(fit)
Random effects:
 Groups Name        Variance Std.Dev. # variability of normal random effects
 person (Intercept) 8.143    2.854
Fixed effects:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -1.8343      0.1624  -11.294    <2e-16
question       0.2100      0.1301    1.614     0.106
---------------------------------------------------------------------------
```

The relatively large value of $\hat{\sigma} = 2.85$ reflects a strong association between the two responses. In fact, Table 10.1 has a sample odds ratio of 10.9. When the sample log odds ratio in such a table is nonnegative, the ML estimate of β for this GLMM is identical to the conditional ML estimate from treating $\{u_i + \alpha\}$ in model (10.2) as fixed effects $\{\alpha_i\}$. Section 8.2.4 presented this conditional ML approach, for which $\hat{\beta} = \log(132/107) = 0.210$.

10.1.4 Differing Effects in GLMMs and Marginal Models

Sections 8.2.3 and 9.1.3 explained that parameters have different interpretations in subject-specific models, such as GLMMs, than in marginal models. The GLMM interpretations are conditional (cluster-specific), given the random effect. By contrast, effects in marginal models are averaged over all clusters (i.e., population-averaged), and so those effects do not refer to a comparison at a fixed value of a random effect.

The marginal model corresponding to subject-specific model (10.2) is

$$\text{logit}[P(Y_1 = 1)] = \alpha + \beta, \quad \text{logit}[P(Y_2 = 1)] = \alpha,$$

where Y_1 is the response about paying higher taxes for a randomly selected subject and Y_2 is the response about cutting living standards for another randomly selected subject. From the marginal counts from Table 10.1, the estimated log odds of a *yes* response was $\hat{\alpha} + \hat{\beta} = \log(359/785) = -0.782$ for paying higher taxes and $\hat{\alpha} = \log(334/810) = -0.886$ for cutting living standards. The estimate of β in this marginal model is the difference between these log odds. This is the log odds ratio, $\hat{\beta} = \log[(359 \times 810)/(785 \times 334)] = 0.104$. We used the GEE method to find this estimate in Section 8.2.2.

This estimated effect $\hat{\beta} = 0.104$ for the marginal model has the same sign but is weaker in magnitude than the estimated effect $\hat{\beta} = 0.210$ for the random-effects model (10.2). The estimated effect for the random-effects model says that for each subject, the estimated odds of a *yes* response on paying higher taxes are $\exp(0.210) = 1.23$ times the estimated odds of a *yes* response on cutting living standards. The estimated effect for the marginal model says that the estimated odds of a *yes* response on paying higher taxes for a randomly selected subject are $\exp(0.104) = 1.11$ times the estimated odds of a *yes* response on cutting living standards for a different randomly selected subject.

When the link function is nonlinear, such as the logit, the population-averaged effects in marginal models are typically smaller in magnitude than the cluster-specific effects in

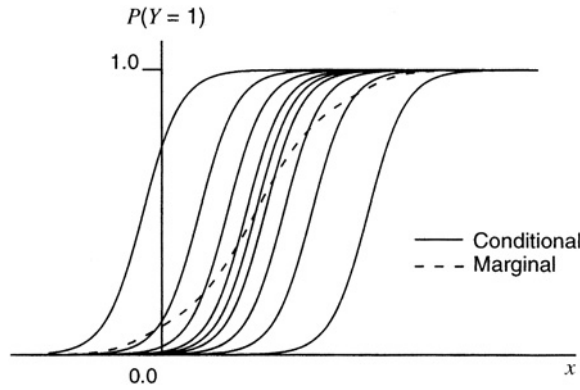


Figure 10.1 Logistic random-intercept model, showing the conditional (subject-specific) curves and the marginal (population-averaged) curve averaging over these.

GLMMs. Figure 10.1 illustrates why. For a single quantitative explanatory variable x, the figure shows cluster-specific logistic regression curves for $P(Y_{it} = 1)$ for several clusters when considerable heterogeneity exists. This corresponds to a relatively large σ for the random effects. At any fixed value of x, variability occurs in the conditional means, $E(Y_{it} \mid u_i) = P(Y_{it} = 1)$. The average of these is the marginal mean, $E(Y_t) = P(Y_t = 1)$. These averages for various x values yield the superimposed dashed curve. That curve shows a weaker effect than each separate curve has. The difference between the two effects is greater as the cluster-specific curves are more spread out, that is, as the variance component σ^2 of the random effects is greater.

10.1.5 Model Fitting for GLMMs

The likelihood function for a GLMM refers to the fixed effects parameters $(\alpha, \beta_1, \ldots, \beta_p)$ and the parameter σ^2 of the $N(0, \sigma^2)$ random effects distribution. To obtain the likelihood function, software eliminates the random effects $\{u_i\}$ by (1) forming the likelihood function as if the $\{u_i\}$ values were known and then (2) averaging that function with respect to the $N(0, \sigma^2)$ distribution of $\{u_i\}$. The main difficulty is step (2), the need to eliminate the random effects in order to obtain the likelihood function. The calculus-based integral used to average with respect to the normal distribution of the random effects does not have a closed form. Numerical methods for approximating it can be computationally intensive.

Gauss–Hermite quadrature is a method that uses a finite sum to approximate the likelihood function. In essence this method approximates the integral in step (2), which is the area under a curve, by the area under a histogram with a particular number of bars. The approximation depends on the number of terms in the finite sum, which is the number q of *quadrature points* at which the function to be integrated is evaluated.[2] As q increases, the estimates and standard errors get closer to the actual ML estimates and their standard errors. Be careful not to use too few quadrature points. Most software will pick a default value for q, often small such as $q = 5$. As a check you can then increase q above that value to make sure the estimates and standard errors have stabilized to the desired degree of

[2] *Adaptive* Gaussian quadrature is a more-refined approximation that more efficiently centers the quadrature points.

precision (e.g., that they do not change in the first few significant digits). The R code shown above in Section 10.1.3 used $q = 50$.

Gauss–Hermite quadrature is feasible when the model contains only a random intercept and possibly a random slope. With a more complex random effect structure, other approaches are needed. *Monte Carlo* methods simulate in order to approximate the relevant integral. The *Laplace approximation* replaces the function to be integrated by an approximation for which the integral has a closed form. Another approach to fitting GLMMs is Bayesian. With it, the distinction between fixed and random effects no longer occurs. A probability distribution (the *prior distribution*) is assumed for each effect of either type.

10.1.6 Inference for Model Parameters and Prediction

For GLMMs, inference about fixed effects proceeds in the usual way. For instance, likelihood-ratio tests can compare models when one model is the special case of the other. Inference about the random effects focuses on their variance σ^2. Predictions $\{\hat{u}_i\}$ for their values are also useful, using the estimated mean of the conditional distribution of u_i, given the data. This prediction depends on *all* the data, not just the data for cluster i.

A significance test for random effects tests H_0: $\sigma = 0$ against H_a: $\sigma > 0$. Since σ cannot be negative, the H_0 case falls on the boundary of the model's parameter space. Because of this, the usual tests are not valid. For example, a Wald statistic such as $\hat{\sigma}/SE$ does not have an approximate standard normal null distribution. (When $\sigma = 0$, an ML estimate $\hat{\sigma} < 0$ is impossible, and $\hat{\sigma}$ is not approximately normally distributed around σ.) The large-sample distribution of the likelihood-ratio statistic, which equals the difference in residual deviances, has probability $1/2$ at 0 and $1/2$ following the shape of a chi-squared distribution with $df = 1$. The test statistic value of 0 occurs when $\hat{\sigma} = 0$, in which case the maximum of the likelihood function is identical under H_0 and H_a. When $\hat{\sigma} > 0$ and the likelihood-ratio test statistic equals t, the P-value is half the right-tail probability above t for a chi-squared distribution with $df = 1$. We will see an example at the end of Section 10.2.2.

10.2 EXAMPLES: RANDOM EFFECTS MODELS FOR BINARY DATA

This section presents examples of GLMMs for binary responses. These are special cases of the logistic-normal model. Analogous models are possible with other link functions, such as the probit.

10.2.1 Small-Area Estimation of Binomial Probabilities

Small-area estimation refers to estimating parameters for many geographical areas that may each have relatively few observations. For example, a study might find county-specific estimates of characteristics such as the unemployment rate or the proportion of families having health insurance coverage. With a national or statewide survey, counties with small populations may have few observations.

Let y_{it} be observation t in area i and let $\pi_i = P(Y_{it} = 1)$ denote the population proportion of *successes* in area i, $i = 1, ..., n$. Let T_i denote the number of observations from area i, of which $y_i = \sum_t y_{it}$ are successes. The fixed effects model

$$\text{logit}[P(Y_{it} = 1)] = \text{logit}(\pi_i) = \alpha_i, \quad i = 1, ..., n,$$

treats the areas as levels of a single factor. When we treat $\{Y_i\}$ as independent binomial random variables, the sample proportions $\{p_i = y_i/T_i\}$ are the ML estimates of $\{\pi_i\}$.

When $\{T_i\}$ are small, sample proportions have large standard errors and may poorly estimate $\{\pi_i\}$. A random effects model that treats each area as a cluster can provide improved estimators. The model is

$$\text{logit}[P(Y_{it} = 1)] = \text{logit}(\pi_i) = u_i + \alpha, \tag{10.3}$$

where $\{u_i\}$ are independent from $N(0, \sigma^2)$. The model now has two parameters (α and σ^2) instead of n parameters. In assuming that the logits of the probabilities vary according to a normal distribution, the fitting process "borrows from the whole," using data from all the areas to estimate the probability in any given one. The estimate for any one area is then a weighted average of the sample proportion for that area alone and the sample proportion after pooling all n samples. The weight given to the sample proportion increases as $\{T_i\}$ increase. As each sample has more data, we put more trust in the separate sample proportions.

Software provides ML estimates $\hat{\alpha}$ and $\hat{\sigma}^2$ and predicted values $\{\hat{u}_i\}$ for the random effects. The estimate of the probability π_i in area i is then

$$\hat{\pi}_i = \exp(\hat{u}_i + \hat{\alpha})/[1 + \exp(\hat{u}_i + \hat{\alpha})].$$

The $\{\hat{\pi}_i\}$ result from shrinking the sample proportions $\{p_i = y_i/T_i\}$ toward the overall sample proportion. The amount of shrinkage decreases as $\{T_i\}$ increase and as $\hat{\sigma}^2$ increases. If $\hat{\sigma}^2 = 0$, then the greatest shrinkage occurs, with $\{\hat{\pi}_i\}$ being identical and equal to the overall sample proportion. When truly all $\{\pi_i\}$ have similar values, $\hat{\pi}_i$ tends to be a much better estimator than the sample proportion p_i from sample i alone.

The simple random effects model (10.3), which is natural for small-area estimation, is useful for any application that estimates many binomial parameters of a similar type when the sample sizes are small. The following example illustrates this.

10.2.2 Example: Estimating Basketball Free Throw Success

In basketball, the player at the center position is usually a tall person who plays mainly close to the basket and may not shoot especially well when not near it. Table 10.2 shows results of "free throws," a standardized shot taken 15 feet from the basket, for the 20 top-scoring centers in the National Basketball Association after one week of the 2016–2017 season.

Let π_i denote the probability that player i makes a free throw, $i = 1, ..., 20$. For T_i observations of player i, we treat the number of successes Y_i as a binomial random variable with index T_i and parameter π_i. For the ML fit of GLMM (10.3), the following R output reports $\hat{\alpha} = 1.174$ for the fixed effect and $\hat{\sigma} = 0.425$ for the random effects:

```
------------------------------------------------------------------------------
> FreeThrow <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                         FreeThrow.dat", header=TRUE) # at text website
> FreeThrow
          player  y  T # player made y free throws in T attempts
1        Davis.A 32 39
...
20      Gobert.R 11 14
```

```
> library(lme4)
> fit <- glmer(y/T ~ 1 + (1|player), family=binomial, weights=T, nAGQ=100,
+              data=FreeThrow) # (1|player) = random intercept for each player
>              # nAGQ = number of points for adaptive Gaussian quadrature
> summary(fit)
     deviance  df.resid
         29.8        18
Random effects:
  Groups Name         Variance  Std.Dev.
  player (Intercept)  0.1807    0.4251
Fixed effects:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)    1.174       0.181    6.484  8.95e-11

> fitted(fit) # estimated prob's for 20 players using predicted random effects
      1       2       3       4       5       6       7       8
0.79485 0.77097 0.65920 0.76541 0.79291 0.79216 0.72353 0.78294
      9      10      11      12      13      14      15      16
0.71108 0.66647 0.80878 0.75504 0.80057 0.80532 0.73415 0.80878
     17      18      19      20
0.74760 0.73415 0.77261 0.77062
-------------------------------------------------------------------------------
```

Table 10.2 Estimates of probability of making a free throw, based on data for centers from week 1 of an NBA season.

Player	T_i	p_i	$\hat{\pi}_i$	π_i	Player	T_i	p_i	$\hat{\pi}_i$	π_i
Davis	39	0.82	0.79	0.80	Lopez	13	0.92	0.81	0.81
Cousins	49	0.78	0.77	0.77	Jokic	3	0.67	0.76	0.82
Whiteside	21	0.52	0.66	0.63	Dieng	6	1.00	0.80	0.81
Turner	13	0.77	0.77	0.81	Adams	7	1.00	0.81	0.61
Gasol	19	0.84	0.79	0.84	Kanter	8	0.62	0.73	0.79
Valanciunas	14	0.86	0.79	0.81	Monroe	13	0.92	0.81	0.74
Towns	3	0.33	0.72	0.83	Horford	6	0.67	0.75	0.80
Embiid	12	0.83	0.78	0.78	Vucevic	8	0.62	0.73	0.67
Nurkic	19	0.63	0.71	0.57	Muscala	10	0.80	0.77	0.77
Drummond	16	0.50	0.67	0.39	Gobert	14	0.79	0.77	0.65

Note: T_i = number of free throws, p_i = sample success proportion, $\hat{\pi}_i$ = random-effects model estimate, π_i = end-of-year success proportion.
Source: `stats.nba.com`, data in `FreeThrow` data file at text website.

For a player with $\hat{u}_i = 0$, the estimated probability of making a free throw is $\exp(1.174)/[1 + \exp(1.174)] = 0.764$. The predicted random effect values yield probability estimates $\{\hat{\pi}_i\}$, also shown in Table 10.2 and in the R output. Since $\{T_i\}$ and $\hat{\sigma}$ are relatively small, these estimates shrink the sample proportions substantially toward the overall sample proportion of successes, which was $222/293 = 0.758$. The $\{\hat{\pi}_i\}$ vary only between 0.66 and 0.81, whereas the sample proportions $\{p_i\}$ vary between 0.33 and 1.0. Relatively extreme sample proportions based on few observations, such as the sample proportion of

0.33 for Towns, shrink more. Table 10.2 also shows the success proportions $\{\pi_i\}$ at the end of the season, which we regard as the parameters estimated by $\{p_i\}$ and $\{\hat{\pi}_i\}$. The average absolute distance from these is 0.124 for the sample proportions and 0.065 for the model-fitted proportions.

When $\sigma^2 = 0$ for the $N(0, \sigma^2)$ distribution of $\{u_i\}$, the model simplifies to

$$\text{logit}(\pi_i) = \alpha,$$

in which the success probability is the same for each player. Here is its fit:

```
------------------------------------------------------------------------------
> summary(glm(y/T ~ 1, family=binomial, weights=T, data=FreeThrow))
            Estimate  Std. Error  z value  Pr(>|z|) # identical player prob's
(Intercept)   1.1400      0.1363    8.361    <2e-16 # null model
---
Residual deviance: 31.322  on 19  degrees of freedom
------------------------------------------------------------------------------
```

The fit has $\hat{\alpha} = 1.140$, for which $e^{1.140}/(1 + e^{1.140}) = 0.758$ is the overall sample proportion of successes. The likelihood-ratio test statistic (difference of deviances) comparing the two models equals 1.56. The P-value is half the right-tail probability above 1.56 for a chi-squared distribution with $df = 1$, which is 0.21. It is plausible that all players have the same success probability. However, $\{T_i\}$ were very small, which is why an implausibly simplistic model seems adequate.

10.2.3 Example: Opinions about Legalized Abortion Revisited

In Sections 9.1.2 and 9.2.3 we analyzed data shown in Section 9.1.2 on opinions about legalized abortion in three situations. Let y_{it} denote the response for subject i in situation t, with $y_{it} = 1$ representing support for legalization. A random intercept model with main effects $\{\beta_t\}$ for the situations factor and γ for gender is

$$\text{logit}[P(Y_{it} = 1)] = u_i + \alpha + \beta_t + \gamma x_i,$$

where $x_i = 1$ for females and 0 for males and $\{u_i\}$ are independent from a $N(0, \sigma^2)$ distribution. In the following R code, we impose the identifiability constraint $\beta_3 = 0$. Then, the parameters compare situations 1 and 2 to situation 3, which states that a woman should be able to get an abortion for any reason.

```
------------------------------------------------------------------------------
>  Abortion <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                          Abortion.dat", header=TRUE)
> Abortion
 person  gender  situation  response
      1       1          1         1
      1       1          2         1
      1       1          3         1
 ...
   1850       0          1         0
```

```
   1850       0          2          0
   1850       0          3          0
> sit <- factor(Abortion$situation, levels=c(3,1,2))
> library(lme4)
> fit <- glmer(response ~ (1|person) + sit + gender, family=binomial,
+              nAGQ=100, data=Abortion)
> summary(fit)
Random effects:
 Groups Name         Variance  Std.Dev.
 person (Intercept)  76.49     8.746
Fixed effects:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -0.61936     0.37820   -1.638     0.101
sit1          0.83478     0.16004    5.216  1.83e-07
sit2          0.29245     0.15670    1.866     0.062 .
gender        0.01261     0.48955    0.026     0.979
--------------------------------------------------------------------------------
```

The fixed effects estimates have subject-specific interpretations. For a subject of either gender, for instance, the estimated odds of supporting legalized abortion in situation 1 equal $\exp(0.835) = 2.30$ times the estimated odds in situation 3. Since $\hat{\gamma} = 0.013$, in each situation the estimated probability of supporting legalization is similar for females and males with identical random effect values. The random effects have $\hat{\sigma} = 8.75$. This is extremely high. People seem to be highly heterogeneous in their response probabilities in any particular situation. The large $\hat{\sigma}$ reflects strong associations among responses in the three sitations. In fact, 1595 of the 1850 subjects made the same response in all three situations. In the US, people tend to be either uniformly opposed to legalized abortion, regardless of the situation, or uniformly in favor of it.

A marginal model analog of this GLMM is

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta_t + \gamma x,$$

where Y_t is the response in situation t for a randomly selected subject. The population-averaged GEE estimates $\{\hat{\beta}_t\}$ for the exchangeable working correlation structure, shown in Section 9.3.1 and again here in Table 10.3, are much smaller than the subject-specific $\{\hat{\beta}_t\}$ from the GLMM. The difference reflects the phenomenon illustrated in Figure 10.1, due to the very large GLMM heterogeneity ($\hat{\sigma} = 8.75$) and the corresponding strong correlations

Table 10.3 Summary of ML estimates for random effects model and GEE estimates for corresponding marginal model with exchangeable working correlation matrix.

		GLMM ML		Marginal Model GEE	
Effect	Parameter	Estimate	*SE*	Estimate	*SE*
Situation	$\beta_1 - \beta_3$	0.835	0.160	0.149	0.030
	$\beta_1 - \beta_2$	0.542	0.157	0.097	0.028
	$\beta_2 - \beta_3$	0.292	0.157	0.052	0.027
Gender	γ	0.013	0.490	0.003	0.088
$\sqrt{\text{Var}(u_i)}$	σ	8.75	0.54		

among the three responses. For instance, the GEE analysis estimates a common correlation of 0.817 between pairs of responses. Although the GLMM $\{\hat{\beta}_t\}$ are about 5–6 times the marginal model $\{\hat{\beta}_t\}$, so are the standard errors. The two approaches provide similar substantive interpretations and conclusions.

10.2.4 Item Response Models: The Rasch Model

In the example just considered comparing opinions in three situations, we have seen that a GLMM without a gender effect,

$$\text{logit}[P(Y_{it} = 1)] = u_i + \alpha + \beta_t,$$

is adequate. Early applications of this form of GLMM were in psychometrics and in educational testing, to describe responses to a battery of T questions on an exam. The probability $P(Y_{it} = 1)$ that subject i makes the correct response on question t depends on the overall ability of subject i, characterized by u_i, and the easiness of question t, characterized by β_t. Such models are called *item response models*.

This logit-link version of the item response model is often called the *Rasch model*, named after a Danish statistician who proposed it for such applications in 1961. Rasch treated the subject terms as fixed effects and used conditional ML methods. These days it is more common to treat subject terms as random effects.

10.2.5 Choice of Marginal Model or Random Effects Model

In practice, how do you decide whether to use a random effects model or instead a marginal model? There is no reason you cannot use both, but each model type has advantages and disadvantages.

With the GEE approach to fitting marginal models, a drawback is that likelihood-based inferences are not available. The GEE approach does not specify a joint distribution of the responses, so it does not have a likelihood function. In addition, a marginal model focuses only on the marginal distributions and does not explicitly include or permit estimating subject-specific effects. Random effects models more fully describe the structure of the data. This approach is preferable if you want to fully model the joint distribution or specify a mechanism that could generate positive association among clustered observations. Latent variable constructions used to motivate model forms usually apply more naturally at the cluster level than the marginal level. The random effects modeling approach is also preferable if a goal is to estimate cluster-specific terms (such as in the example of free-throw success) or estimate their variability or have fixed effects refer to the cluster level. For example, some methodologists use random effects models whenever their main focus is on "within-cluster" effects.

By contrast, if the main focus is on comparing groups that are independent samples, effects of interest are "between-cluster" rather than within-cluster. In many surveys and epidemiological studies, a goal is to compare the relative frequency of some outcome for different groups in a population. Then, quantities of primary interest include between-group comparisons of marginal probabilities for the different groups. For instance, for the opinions about legalized abortion example, if a study focused on estimating the probability of supporting legalization and comparing opinions of females and males, then marginal probabilities are mainly relevant. When marginal effects are the main focus, it is simpler to use a

marginal model to model the margins and directly estimate those effects. Developing a more detailed model of the joint distribution that generates those margins, as a random effects model does, provides greater opportunity for misspecification. For instance, with longitudinal data the assumption that observations are independent, given the random effect, or that different groups have the same variability of the random effects, need not be realistic.

10.3 EXTENSIONS TO MULTINOMIAL RESPONSES AND MULTIPLE RANDOM EFFECT TERMS

GLMMs extend directly from binary outcomes to multicategory outcomes. Modeling is simpler with ordinal responses than with nominal responses, because it is usually adequate to use the same random effect term for each logit. With cumulative logits, this is the *proportional odds* structure that Section 6.2.1 used for fixed effects. However, GLMMs can have more than one random effect term in a model, such as to allow random slopes as well as random intercepts. This section shows examples of these two extensions.

10.3.1 Example: Insomnia Study Revisited

Table 9.2 showed results of a clinical trial comparing an active hypnotic drug with placebo on two occasions in treating insomnia patients. The response, time to fall asleep, falls in one of four ordered categories. We analyzed the data with a marginal model in Section 9.3.1 and with a transitional model in Section 9.4.3. Let y_{it} denote the response for subject i at time t (0 = initial, 1 = follow-up) using treatment x (1 = active drug, 0 = placebo).

Using GEE, in Section 9.3.1 we fitted the marginal model

$$\text{logit}[P(Y_{it} \leq j)] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3 (t \times x).$$

The corresponding cumulative logit random-intercept model is

$$\text{logit}[P(Y_{it} \leq j)] = u_i + \alpha_j + \beta_1 t + \beta_2 x + \beta_3 (t \times x).$$

Some software, such as shown next with R, fits the model with the alternative parameterization

$$\text{logit}[P(Y_{it} \leq j)] = u_i + \alpha_j - \beta_1 t - \beta_2 x - \beta_3 (t \times x),$$

implied by the latent-variable model introduced in Section 6.2.6. This changes the signs of the effect estimates, but not their standard errors or inference results. Here are R results:

```
-------------------------------------------------------------------------------
> Insomnia <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Insomnia.dat",
+                        header=TRUE)
> Insomnia
    case  treat occasion  response
1      1      1        0         1
2      1      1        1         1
...
477  239      0        0         4
478  239      0        1         4
```

```
> library(ordinal)
> fit <- clmm(factor(response) ~ (1|case) + occasion + treat + occasion:treat,
+             nAGQ=20, data=Insomnia) # response var. for clmm must be factor
> summary(fit)
Random effects:
 Groups Name         Variance  Std.Dev.
 case   (Intercept)  3.628     1.905
Coefficients: # not showing intercept estimates
               Estimate  Std. Error  z value  Pr(>|z|)
occasion       -1.60158     0.28336   -5.652  1.58e-08 # estimates have
treat          -0.05785     0.36629   -0.158   0.87450 # opposite sign for
occasion:treat -1.08129     0.38046   -2.842   0.00448 # +beta parameters
-------------------------------------------------------------------------------
```

The model fit suggests that (1) since $\hat{\beta}_2$ is close to 0, response distributions are similar initially for the two treatments; (2) since $\hat{\beta}_1$ is significantly negative, the placebo group tends to fall asleep more quickly at the follow-up time than initially; (3) since the interaction term $\hat{\beta}_3$ is significantly negative, the time effect is greater for the active drug than for placebo. As in the marginal model analysis, we conclude that time to fall asleep tends to decrease more for the active drug than for placebo.

Table 10.4 summarizes results for the GLMM and the marginal model. Estimates and standard errors are about 50% larger for the GLMM. This reflects the considerable heterogeneity. The random effects have $\hat{\sigma} = 1.90$, corresponding to a moderate association between the responses at the two occasions. (With scores (10, 25, 45, 75), the correlation between the initial and follow-up responses is 0.44.)

Table 10.4 Results of fitting cumulative logit marginal model and random effects model to Table 9.2, with standard errors in parentheses.

Effect	Marginal Model GEE	Random Effects Model (GLMM) ML
Occasion	1.038 (0.168)	1.602 (0.283)
Treatment	0.034 (0.238)	0.058 (0.366)
Treatment×Occasion	0.708 (0.244)	1.081 (0.380)

10.3.2 Meta-Analysis: Bivariate Random Effects for Association Heterogeneity

The examples so far have used univariate random effects, in the form of random intercepts. Sometimes it is natural to have a random slope as well as a random intercept. We illustrate using Table 10.5, which shows results from 3 of 41 studies that compared a new surgery to an older surgery for treating ulcers. The analyses below use data from all 41 studies, for which data are at the text website. The response was whether the surgery resulted in the adverse event of recurrent bleeding (1 = yes, 0 = no). A statistical analysis that summarizes results from several studies about the same association is called a *meta-analysis*.

As usual, to compare two groups on a binary response variable with data stratified on a third variable, we can analyze the strength of association in the 2×2 tables and investigate how that association varies (if at all) among the strata. When the strata are themselves a sample, such as different studies for a meta analysis, or schools, or medical clinics, a random

Table 10.5 Tables relating surgery treatment (new or old) to outcome on an adverse event, for 3 of the 41 studies.

Study	Surgery Treatment	Adverse Event		Sample Odds Ratio	Model Fitted Odds Ratio
		Yes	No		
1	New	7	8	0.159	0.147
	Old	11	2		
5	New	3	9	∞	2.59
	Old	0	12		
6	New	4	3	0.0	0.126
	Old	4	0		

Source: Article by B. Efron, *J. Amer. Statist. Assoc.* **91**: 539 (1996). Complete data for 41 studies available in `Ulcers` data file at text website.

effects approach is natural. We then use a separate random effect for each stratum rather than for each subject. With a random sampling of strata, we can extend inferences to the population of strata.

Let y_{it} denote the response for a subject in study i using surgery treatment t ($1 =$ new; $2 =$ old). One possible model is the logistic-normal random intercept model,

$$\text{logit}[P(Y_{i1} = 1)] = u_i + \alpha + \beta, \quad \text{logit}[P(Y_{i2} = 1)] = u_i + \alpha.$$

Each study is a cluster having a random effect. This model treats the log odds ratio β between treatment and response as the same in each study. The parameter σ^2 summarizes study-to-study heterogeneity in the logit-probabilities of an adverse event. The estimated treatment effect is $\hat{\beta} = -1.173$ ($SE = 0.118$), as shown in the following `R` output:

```
-------------------------------------------------------------------------------
> Ulcers  <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Ulcers.dat",
+                       header=TRUE)
> Ulcers
   study treat  y  n # y "yes" outcomes on adverse events, sample size n
1      1     1  7 15
2      1     0 11 13
...
81    41     1  0  9
82    41     0  0 16
> library(lme4)
> fit <- glmer(y/n ~ (1|study) + treat, family=binomial, weights=n, nAGQ=50,
+              data=Ulcers)
> summary(fit)
Random effects:
 Groups Name         Variance  Std.Dev.
 study  (Intercept)  0.6721    0.8198
Fixed effects:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -0.3212      0.1500   -2.142    0.0322
treat         -1.1728      0.1176   -9.969    <2e-16
-------------------------------------------------------------------------------
```

This is similar to the estimated treatment effect from treating the study terms $\{u_i\}$ as fixed rather than random ($\hat{\beta} = -1.220$, $SE = 0.119$). The evidence is strong that adverse events are less likely with the new surgery.

It is more realistic to allow the treatment effect to vary across the 41 studies. A logistic-normal model permitting treatment-by-study interaction is

$$\begin{aligned} \text{logit}[P(Y_{i1} = 1)] &= u_i + \alpha + (\beta + v_i), \\ \text{logit}[P(Y_{i2} = 1)] &= u_i + \alpha. \end{aligned} \tag{10.4}$$

Here, u_i is a random intercept and v_i is a random slope in the sense that it is the coefficient of an indicator variable for surgery treatment (1 = new, 0 = old). The log odds ratio between treatment and response equals $\beta + v_i$ in study i. We assume that $\{(u_i, v_i)\}$ have a *bivariate normal* distribution. That distribution has means 0 and variances σ_u^2 for $\{u_i\}$ and σ_v^2 for $\{v_i\}$, with a correlation ρ between u_i and v_i, $i = 1, \ldots, 41$. Therefore, β is the mean study-specific log odds ratio and σ_v describes variability in the log odds ratios. The fit of the model provides a simple summary of an estimated mean $\hat{\beta}$ and an estimated standard deviation $\hat{\sigma}_v$ of the log odds ratios for the population of strata.

The sample log odds ratios vary considerably among the 41 studies. Some sample odds ratios even take the boundary values of 0 or ∞, as Table 10.5 shows. The following R output shows the fit of model (10.4):

```
------------------------------------------------------------------------------
> fit2 <- glmer(y/n ~ treat + (1+treat|study), family=binomial, weights=n,
+               data=Ulcers)
> summary(fit2)
Generalized linear mixed model fit by maximum likelihood (Laplace Approx.)
Random effects:
 Groups Name         Variance  Std.Dev. Corr
 study  (Intercept)  2.015     1.420
        treat        2.242     1.497   -0.87
Fixed effects:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -0.2415      0.2416    -1.00     0.317
treat         -1.2965      0.2724    -4.76  1.94e-06
> fitted(fit2) # predicted proportions for the 82 study-treatment cases
      1        2        3        4        5        6
0.36219  0.79493  0.33822  0.50038  0.15019  0.12413
...
      82
0.068815
------------------------------------------------------------------------------
```

The estimated mean log odds ratio is[3] $\hat{\beta} = -1.2965$ ($SE = 0.272$), which corresponds to a summary odds ratio estimate of 0.27. Considerable heterogeneity seems to occur in the true log odds ratios, suggested by $\hat{\sigma}_v = 1.50$.

[3] The *glmer* R function used here merely approximates ML results by the *Laplace approximation*. The actual ML results are $\hat{\beta} = -1.299$ ($SE = 0.277$) and $\hat{\sigma}_v = 1.52$.

The hypothesis H_0: $\beta = 0$ specifies an absence of association between treatment and response, in the sense of a mean log odds ratio of 0. The Wald statistic $z = \hat{\beta}/SE = -1.2965/0.272 = -4.76$ shows strong evidence against H_0. However, the evidence is weaker than for the model without treatment-by-study interaction, for which $z = \hat{\beta}/SE = -1.173/0.118 = -9.97$. The extra variance component in the interaction model pertains to variability in the log odds ratios. As its estimate $\hat{\sigma}_v$ increases, so does the standard error of the estimated treatment effect $\hat{\beta}$ tend to increase. The more that the treatment effect varies among studies, the more difficult it is to estimate precisely the mean of that effect.

Table 10.5 shows sample odds ratios and the model fitted odds ratios for three studies. For Study 1, for instance, the R output shows predicted probabilities of an adverse event of 0.362 for the new surgery and 0.795 for the old surgery, yielding a predicted odds ratio of 0.147. For all 41 studies, the sample odds ratios vary from 0.0 to ∞. Their random effects model counterparts vary only between 0.004 (for a study that reported 0 out of 34 adverse events for the new surgery and 34 out of 34 adverse events for the old surgery!) and 2.6 (for study 5). The model-based estimates are much less variable. They do not have the same ordering as the sample values, because the shrinkage tends to be greater for studies having smaller sample sizes.

10.4 MULTILEVEL (HIERARCHICAL) MODELS

Hierarchical models describe observations that have a nested nature: units at one level are contained within units of another level. Models having a hierarchical structure are called *multilevel models*. Such models usually treat terms for the units as random effects rather than fixed effects, especially when those units are regarded as sampled from a population of interest. The multilevel model then contains random effect terms for the different levels of units.

10.4.1 Example: Two-Level Model for Student Performance

Hierarchical data are common in certain application areas, such as in educational studies. A study of factors that affect student performance might measure, for each student and each exam in a battery of exams, whether the student passed. Students are nested within schools, and the model could incorporate variability among students as well as variability among schools. The model could analyze effects of characteristics of the student, such as $x_1 =$ gender and $x_2 =$ score on each achievement exam a year ago, and effects of characteristics of the school the student attends, such as $x_3 =$ the school budget, per student, and $x_4 =$ average class size. Observations for the same student on different exams would probably tend to be more alike than observations for different students. Likewise, students in the same school might tend to have more-alike observations than students from different schools, because students within a school tend to be similar on characteristics such as socioeconomic status. Relevant models contain explanatory variables and random effect terms for the student and for the school.

Let $y_{ijt} =$ whether student i in school j passed exam t (1 = yes, 0 = no). For the student-specific explanatory variables just mentioned, the level-one model has form

$$\text{logit}[P(Y_{ijt} = 1)] = u_{ij} + \alpha_j + \beta_1 x_{1ij} + \beta_2 x_{2ijt}.$$

The random effect u_{ij} for student i in school j accounts for variability among students in student-specific explanatory variables not measured, such as perhaps student ability and mother's and father's attained educational level. For the school-specific explanatory variables, the level-two model takes the school-specific term α_j from the level-one model and expresses it as

$$\alpha_j = s_j + \alpha + \beta_3 x_{3j} + \beta_4 x_{4j},$$

where s_j is a random effect for school j. This random effect reflects heterogeneity among the schools due to school-specific explanatory variables not measured, such as perhaps the teachers' average salary, the degree of drug-related problems in the school, and characteristics of the district for which the school enrolls students. Substituting the level-two model into the level-one model, we obtain

$$\text{logit}[P(Y_{ijt} = 1)] = u_{ij} + s_j + \alpha + \beta_1 x_{1ij} + \beta_2 x_{2ijt} + \beta_3 x_{3j} + \beta_4 x_{4j}.$$

This is a multilevel model with random intercepts u_{ij} at the student level and s_j at the school level.

10.4.2 Example: Smoking Prevention and Cessation Study

We illustrate multilevel models using a study[4] of the efficacy of two programs for discouraging young people from starting or continuing to smoke. The study compared four groups, defined by a 2×2 factorial design according to whether a student was exposed to a school-based curriculum (SC; 1 = yes, 0 = no) and a television-based prevention program (TV; 1 = yes, 0 = no). The subjects were 1600 seventh-grade students from 135 classrooms in 28 Los Angeles schools. The schools were randomly assigned to the four intervention conditions. The response variable was a tobacco and health knowledge (THK) scale, measured at the end of the study. This variable was also observed at the beginning of the study, and that measure (PTHK = Pre-THK) was used as a covariate. THK took values between 0 and 7, with $\bar{y} = 2.66$ and $s_y = 1.38$. The data are shown partly in Table 10.6.

Table 10.6 Part of smoking prevention and cessation data file.[a]

Student	School	Class	SC	TV	PTHK	THK
1	403	403101	1	0	2	3
2	403	403101	1	0	4	4
...						
1600	515	515113	0	0	3	3

[a]Complete `Smoking` data file, courtesy of Don Hedeker, is at the text website.

Let y_{ijk} denote whether the follow-up THK score for student i within classroom j in school k exceeds 2 (1 = yes, 0 = no). The multilevel model

$$\text{logit}[P(Y_{ijk} = 1)] = c_{jk} + s_k + \alpha + \beta_1 \text{PTHK}_{ijk} + \beta_2 \text{SC}_{ijk} + \beta_3 \text{TV}_{ijk}$$

[4] See p. 9 of *Longitudinal Data Analysis* by D. Hedeker and R. Gibbons (2006, Wiley).

has random effects for classrooms and schools. At one level, the random effect c_{jk} for classroom j in school k is assumed to have a $N(0, \sigma_c^2)$ distribution, with unknown σ_c^2. At another level, the random effect s_k for school k is assumed to have a $N(0, \sigma_s^2)$ distribution, with unknown σ_s^2. Within a particular classroom and school, the responses for students in that classroom and school are assumed to be independent. Each student has only a single observation on the response variable, so the model does not contain a student random effect.

Of the factorial fixed effects, the following R output shows that SC is highly significant but TV is not significant:

```
------------------------------------------------------------------------------
> Smoking <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Smoking.dat",
+                       header=TRUE)
> Smoking
     student school  class SC TV PTHK THK y
1          1    403 403101  1  0    2   3 1
...
1600    1600    515 515113  0  0    3   3 1
> library(lme4)
> fit <- glmer(y ~ (1|class) + (1|school) + PTHK + SC + TV, family=binomial,
               data=Smoking)
> summary(fit) # Gaussian quadrature not available for multilevel models
GLMM fit by maximum likelihood (Laplace Approximation)
Random effects:
 Groups Name         Variance  Std.Dev.
 class  (Intercept)  0.16728   0.4090
 school (Intercept)  0.06413   0.2532
Number of obs: 1600, groups: class, 135; school, 28
Fixed effects:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -1.13163     0.17827   -6.348  2.18e-10
PTHK         0.39512     0.04627    8.539   < 2e-16
SC           0.80014     0.16893    4.737  2.17e-06
TV           0.10786     0.16819    0.641     0.521
------------------------------------------------------------------------------
```

The estimated standard deviations of the random effects, $\hat{\sigma}_c = 0.41$ and $\hat{\sigma}_s = 0.25$, indicate slightly more variability among classrooms within schools than among schools. Adding an interaction between SC and TV (not shown here) does not significantly improve the fit.

Suppose we had ignored the clustering of observations in classrooms and schools and treated the 1600 observations as independent by fitting the ordinary logistic regression model,

$$\text{logit}[P(Y_{ijk} = 1)] = \alpha + \beta_1 \text{PTHK}_{ijk} + \beta_2 \text{SC}_{ijk} + \beta_3 \text{TV}_{ijk}.$$

Would it make a difference? Edited R output follows:

```
------------------------------------------------------------------------------
> fit.glm <- glm(y ~ PTHK + SC + TV, family=binomial, data=Smoking)
> summary(fit.glm) # ignoring hierarchical clustering (no random effects)
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -1.11920     0.13127   -8.526   < 2e-16
PTHK         0.39886     0.04401    9.063   < 2e-16
SC           0.76532     0.10563    7.245  4.32e-13
TV           0.12305     0.10462    1.176      0.24
------------------------------------------------------------------------------
```

The estimated fixed effects are similar to those in the multilevel model, but the standard errors are substantially underestimated for the between-subjects effects (SC and TV).

For more detailed introductions to multilevel modeling, see Bartholomew et al. (2008, Chapter 12), Gelman and Hill (2006), Hedeker and Gibbons (2006), and Snijders and Bosker (2011).

10.5 LATENT CLASS MODELS *

Ordinary GLMMs create a mixture of linear predictor values using an unobserved random effect, which is a latent variable assumed to have a normal distribution. By contrast, latent class models use a mixture distribution that is qualitative rather than quantitative. The basic model assumes existence of a latent categorical variable such that a set of response variables are conditionally independent, given that variable.

10.5.1 Independence Given a Latent Categorical Variable

For T categorical response variables $(Y_1, Y_2, \ldots, Y_T)$, the latent class model without explanatory variables assumes a latent categorical variable Z such that for each possible sequence of response outcomes $(y_1, \ldots, y_T)$ and each category z of Z,

$$P(Y_1 = y_1, \ldots, Y_T = y_T \mid Z = z) = P(Y_1 = y_1 \mid Z = z) \cdots P(Y_T = y_T \mid Z = z).$$

Figure 10.2 shows the independence graph. A latent class model summarizes probabilities of classification $P(Z = z)$ in the latent classes as well as conditional probabilities $P(Y_t = y_t \mid Z = z)$ of outcomes for each Y_t within each latent class. These are the model parameters.[5]

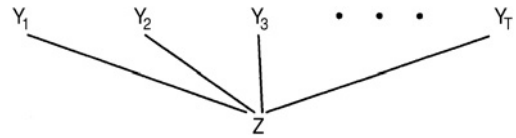


Figure 10.2 Independence graph showing conditional independences for the latent class model.

[5] The model is an analog for categorical responses of the *factor analysis* model with a common factor for multivariate normal response variables.

A latent class model is often plausible when the observed variables are several indicators of some concept, such as prejudice, religiosity, or opinion about an issue. An example is Table 9.1, in which subjects gave their opinions about whether abortion should be legal in various situations. Perhaps an underlying latent variable describes one's basic attitude toward legalized abortion, such that given the value of that latent variable, responses on the observed variables are conditionally independent. For instance, there may be three latent classes: one for those who tend to oppose legalized abortion regardless of the situation, one for those who tend to support it, and one for those whose response tends to change depending on the situation.

A T-dimensional contingency table cross-classifies $(Y_1, \ldots, Y_T)$. The $(T+1)$-dimensional table that cross-classifies the T response variables with the latent variable Z is an unobserved table. Denote the number of categories of each Y_t by c and the number of latent classes of Z by q. The latent class model assumes a multinomial distribution over the c^T cells for the observed data. Its conditional independence factorization states that those joint multinomial cell probabilities satisfy

$$P(Y_1 = y_1, \ldots, Y_T = y_T) = \sum_{z=1}^{q} \left[\prod_{t=1}^{T} P(Y_t = y_t | Z = z) \right] P(Z = z).$$

The latent class model is equivalent to the loglinear model symbolized by $(Y_1Z, Y_2Z, \ldots, Y_TZ)$ for the unobserved table. The model makes no assumption about the $\{Y_tZ\}$ associations but assumes that the $\{Y_t\}$ are mutually independent within each category of Z.

Fitting latent class models requires iterative methods such as the Newton–Raphson algorithm. Often, the nature of the variables suggests a value for q, usually quite small (2 to 4). Otherwise, you can start with $q = 2$, and if the fit is inadequate, increase q by steps of 1 as long as the fit shows substantive improvement. As q increases, however, the danger increases that the likelihood function has multiple peaks, in which case the iterative algorithm may converge to a local rather than global maximum. To lessen the chance that this happens, you should run the latent class software function multiple times with different starting values.

10.5.2 Example: Latent Class Model for Rater Agreement

Table 10.7 shows results for seven pathologists who classified each of 118 slides on the presence or absence of carcinoma in the uterine cervix. For modeling interobserver agreement, the conditional independence assumption of the latent class model is often plausible. With a blind rating scheme, ratings of a given subject or unit by different pathologists are independent. If subjects having true rating in a given category are relatively homogeneous, then ratings by different pathologists may be nearly independent within a given true rating class. Thus, one might posit a latent class model with $q = 2$ classes, one for subjects whose true rating is positive and one for subjects whose true rating is negative.

Table 10.8 shows goodness-of-fit results for some latent class models applied as shown below with R to the ungrouped data file. The grouped data are a contingency table that cross-classifies the diagnoses by the $T = 7$ pathologists. It has $2^7 = 128$ cells, of which 108 are empty. Table 10.7 also shows the fitted values for latent class models with $q = 1, 2, 3$ latent

Table 10.7 Diagnoses of carcinoma (1 = present, 0 = absent) and fits of latent class models with q classes.

Pathologist								Model Fitted Values		
A	B	C	D	E	F	G	Count	$q = 1$	$q = 2$	$q = 3$
0	0	0	0	0	0	0	34	1.1	23.0	33.8
0	0	0	0	1	0	0	2	1.6	6.6	2.0
0	1	0	0	0	0	0	6	2.2	12.7	6.3
0	1	0	0	0	0	1	1	2.8	1.7	1.5
0	1	0	0	1	0	0	4	3.3	3.6	3.0
0	1	0	0	1	0	1	5	4.2	0.5	4.7
1	0	0	0	0	0	0	2	1.4	3.0	2.1
1	0	1	0	1	0	1	1	1.6	0.2	0.2
1	1	0	0	0	0	0	2	2.8	1.7	1.3
1	1	0	0	0	0	1	1	3.5	0.3	1.6
1	1	0	0	1	0	0	2	4.2	0.5	2.9
1	1	0	0	1	0	1	7	5.3	3.7	6.5
1	1	0	0	1	1	1	1	1.4	2.6	1.4
1	1	0	1	0	0	1	1	1.3	0.1	0.1
1	1	0	1	1	0	1	2	2.0	4.3	2.6
1	1	0	1	1	1	1	3	0.5	3.1	2.0
1	1	1	0	1	0	1	13	3.3	11.5	9.6
1	1	1	0	1	1	1	5	0.9	8.4	8.7
1	1	1	1	1	0	1	10	1.2	13.5	13.6
1	1	1	1	1	1	1	16	0.3	9.9	12.3

Source: Based on data in the article by J.R. Landis and G.G. Koch, *Biometrics* **33**: 363–374 (1977), not showing empty cells. Ungrouped data are in `Carcinoma` data file at text website. Fits obtained with Latent Gold (Statistical Innovations, Belmont, MA).

classes, for the cells having positive counts. (Each empty cell also has a fitted value, not shown.) Because the table is so highly sparse, the deviance for that table or for the ungrouped data is mainly useful for comparing models. The model with $q = 1$ latent class is the model of mutual independence of the seven ratings. It fits poorly, as we would expect. With $q = 2$, considerable evidence remains of lack of fit. For instance, the fitted count for a negative rating by each pathologist is 23.0, compared with an observed count of 34. The small deviance that Table 10.8 reports for this model does not imply a good fit; for models for categorical data, the deviance for highly sparse grouped data or ungrouped data tends to be highly conservative (i.e., giving chi-squared P-values that are too large). The model with $q = 3$ seems to fit adequately, as does the model with $q = 4$, but AIC favors the simpler model.

Table 10.8 Goodness-of-fit of latent class models fitted to ungrouped carcinoma diagnosis data.

Number of Latent Classes	Model	Residual Deviance	*df*	AIC
1	Mutual independence	476.8	111	1062.9
2	Latent class	62.4	103	664.5
3	Latent class	15.3	95	633.4
4	Latent class	6.4	87	640.6

Studying the estimated probability $P(Y_t = 1 \mid Z = z)$ of a carcinoma diagnosis for each pathologist, conditional on a given latent class z, helps illuminate the nature of these classes. Table 10.9 reports these for the three-class model. They suggest that (1) the first latent class refers to cases in which A, B, E, and G agree show carcinoma and C and D usually agree; (2) the second latent class refers to cases of strong disagreement, whereby C, D, and F rarely diagnose carcinoma but B, E, and G usually do; and (3) the third latent class refers to cases that all pathologists usually agree show no carcinoma. The estimated proportions in the three latent classes are $\hat{P}(Z = 1) = 0.445$, $\hat{P}(Z = 2) = 0.182$, and $\hat{P}(Z = 3) = 0.374$. The model estimates that 18.2% of the cases fall in the problematic disagreement class.

Table 10.9 Estimated conditional probabilities of diagnosing carcinoma, conditional on the latent class, for model with three latent classes.

Latent	Pathologist						
Class	A	B	C	D	E	F	G
1	1.000	0.981	0.858	0.586	1.000	0.476	1.000
2	0.513	1.000	0.000	0.058	0.751	0.000	0.631
3	0.057	0.138	0.000	0.000	0.055	0.000	0.000

Some R code follows for fitting latent class models and obtaining such conditional probability estimates:

```
------------------------------------------------------------------------------
> Carcinoma <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                         Carcinoma.dat", header=TRUE)
> Carcinoma <- -Carcinoma + 2 # poLCA requires response values 1, 2
> Carcinoma
    A B C D E F G
1   2 2 2 2 2 2 2 # 1 = carcinoma present, 2 = carcinoma absent
...
118 1 1 1 1 1 1 1 # 118 observations in ungrouped data file
> library(poLCA)
> poLCA(cbind(A,B,C,D,E,F,G) ~ 1, nclass=1, data=Carcinoma)
residual degrees of freedom: 111
AIC(1): 1062.93
G^2(1): 476.7814 (Likelihood ratio/deviance statistic)
> poLCA(cbind(A,B,C,D,E,F,G) ~ 1, nclass=2, nrep=9, data=Carcinoma)
residual degrees of freedom: 103
AIC(2): 664.5137 # nrep=9 requests 9 fits with different starting values
G^2(2): 62.36543 (Likelihood ratio/deviance statistic)
> poLCA(cbind(A,B,C,D,E,F,G) ~ 1, nclass=3, nrep=9, data=Carcinoma)
residual degrees of freedom: 95
AIC(3): 633.41
G^2(3): 15.26171 (Likelihood ratio/deviance statistic)
Conditional item response probabilities, by outcome variable, for each class
$A
           Pr(1)  Pr(2) # 1 = carcinoma present 2 = carcinoma absemt
class 1:  1.0000 0.0000 # estimated P(carcinoma is present) are
```

```
class 2:  0.5128 0.4872 # (1.000, 0.513, 0.057) for latent classes 1, 2, 3
class 3:  0.0573 0.9427
...
$G
           Pr(1)  Pr(2)
class 1:  1.0000 0.0000
class 2:  0.6307 0.3693
class 3:  0.0000 1.0000

Estimated class population shares
 0.4447 0.1817 0.3736
> poLCA(cbind(A,B,C,D,E,F,G) ~ 1, nclass=4, nrep=9, data=Carcinoma)
residual degrees of freedom: 87
AIC(4): 640.5717 # Caution: different starting values gave different results
G^2(4): 6.42345 (Likelihood ratio/deviance statistic)
-------------------------------------------------------------------------------
```

Using model parameter estimates and Bayes' Theorem, we can also estimate $P(Z = z \mid Y_t = y_t)$ and $P(Z = z \mid Y_1 = y_1, \ldots, Y_T = y_T)$. If a pathologist makes a *yes* rating, for instance, what is the estimated probability that the subject is in the latent class for which agreement on a positive rating usually occurs? We could also use a GLMM with a normal rather than categorical latent variable. A logistic-normal random intercept model, for instance, yields case-specific comparisons of $P(Y_t = 1)$ for various t.

A danger with latent variable models, shared by factor analysis for continuous responses, is the temptation to interpret latent variables too literally. For example, here it is tempting to treat latent class 1 as cases truly having carcinoma and a rating of carcinoma given that the subject falls in latent level 1 as being a correct judgment. Realize the tentative nature of the latent variable and be careful not to make the error of reification — treating an abstract construction as if it has actual existence.

The basic latent class model generalizes in many ways. For example, *latent class regression models* can have explanatory variables. The latent variable can be multivariate. *Latent transition models* permit transitioning over time between classes. *Mixed-membership models* permit each subject to have partial membership in various classes.

For further details about latent class models and their generalizations, see Agresti (2013, Chapter 14), Bartholomew et al. (2011), Collins and Lanza (2009), and Magidson and Vermunt (2004). For further details about the methods presented in Chapters 9 and 10 on clustered, categorical data, see Fitzmaurice et al. (2011), Gelman and Hill (2006), Gueorguieva (2018), Hedeker and Gibbons (2006), and Molenberghs and Verbeke (2005).

EXERCISES

10.1 For Table 8.9, which asked subjects whether they believe in heaven and whether they believe in hell, fit model (10.2). If your software uses numerical integration, report $\hat{\beta}$, $\hat{\sigma}$, and their standard errors for $q = 2$, 10, 100, 400, and 1000 quadrature points. Comment on convergence. Interpret $\hat{\beta}$.

10.2 Table 10.10 shows results of a teratology experiment, available in the `Teratology` data file at the text website. Female rats on iron-deficient diets were assigned to

four groups. Group 1 received only placebo injections. The other groups received injections of an iron supplement according to various schedules. The rats were made pregnant and then sacrificed after three weeks. For each fetus in each rat's litter, the response was whether the fetus was dead.

a. Let y_i denote the number of dead fetuses for the T_i fetuses in litter i. Let π_{it} denote the probability of death for fetus t in litter i. Let $z_{ig} = 1$ if litter i is in group g and 0 otherwise. Ignoring the clustering and treating y_i as a binomial variate, fit the model

$$\text{logit}(\pi_{it}) = \alpha + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4}.$$

Allowing observations within each litter to be correlated, use the GEE approach to fit the model. Compare parameter estimates and standard errors. Interpret.

b. Specify the corresponding GLMM. Fit it and compare results to those in (**a**).

Table 10.10 Response counts of (litter size, number dead) for 58 litters of rats in low-iron teratology study.

Group 1: Untreated (Low Iron)
(10, 1) (11, 4) (12, 9) (4, 4) (10, 10) (11, 9) (9, 9) (11, 11) (10, 10) (10, 7) (12, 12) (10, 9) (8, 8) (11, 9) (6, 4) (9, 7) (14, 14) (12, 7) (11, 9) (13, 8) (14, 5) (10, 10) (12, 10) (13, 8) (10, 10) (14, 3) (13, 13) (4, 3) (8, 8) (13, 5) (12, 12)
Group 2: Injections Days 7 and 10
(10, 1) (3, 1) (13, 1) (12, 0) (14, 4) (9, 2) (13, 2) (16, 1) (11, 0) (4, 0) (1, 0) (12, 0)
Group 3: Injections Days 0 and 7
(8, 0) (11, 1) (14, 0) (14, 1) (11, 0)
Group 4: Injections Weekly
(3, 0) (13, 0) (9, 2) (17, 2) (15, 0) (2, 0) (14, 1) (8, 0) (6, 0) (17, 0)

Source: D.F. Moore and A. Tsiatis, *Biometrics* **47**: 383–401 (1991).

10.3 For 10 coins, let π_i denote the probability of a head for coin i. You flip each coin 5 times. The sample numbers of heads are (2, 4, 1, 3, 3, 5, 4, 2, 3, 1).

a. Report the sample proportion estimates of π_i. Formulate a model for which these are the ML estimates.

b. Formulate a random effects model for the data. Using software, obtain predicted values $\{\hat{\pi}_i\}$.

c. Suppose all $\pi_i = 0.50$. Compare the estimates in (**a**) and in (**b**) by finding the average absolute distance of the estimates from 0.50 in each case. What does this suggest?

10.4 For Table 7.1, let $y_{it} = 1$ when student i used substance t ($t = 1$, cigarettes; $t = 2$, alcohol; $t = 3$, marijuana). Fit the model, $\text{logit}[P(Y_{it} = 1)] = u_i + \alpha + \beta_t$.

a. Report and interpret the estimated fixed effects. Conduct a likelihood-ratio test of H_0: $\beta_1 = \beta_2 = \beta_3$ as a way of testing marginal homogenity.

b. Report $\hat{\sigma}$ for the random effects. In practical terms, what does (i) the large value imply? (ii) a large positive value for u_i for a particular student represent?

c. Why are $\{\hat{\beta}_t\}$ so different from $\{\hat{\beta}_t\}$ for the marginal model in Exercise 9.1?

d. Explain how the focus differs for the random effects and marginal models than for the loglinear model (AC, AM, CM) fitted to these data in Section 7.1.7.

e. Adding race and gender to the analysis, with data as shown in Table 7.8, (**i**) analyze using GLMMs, (**ii**) compare results and interpretations to those with marginal models in Exercise 9.3.

10.5 For the crossover study summarized in Exercise 9.6, fit the model

$$\text{logit}[P(Y_{i(k)t} = 1)] = u_{i(k)} + \alpha_k + \beta_t,$$

where $\{u_{i(k)}\}$ are independent $N(0, \sigma^2)$. Interpret $\{\hat{\beta}_t\}$ and $\hat{\sigma}$. Compare interpretations and estimates of $\beta_B - \beta_A$ and $\beta_C - \beta_A$ and their SE values to those using the marginal model of Exercise 9.6.

10.6 Refer to Exercise 9.4 and the `Depression` data file. Now let y_{it} denote observation t for subject i. Fit the model

$$\text{logit}[P(Y_{it} = 1)] = u_i + \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 (d \times t).$$

Interpret effects and compare to the corresponding marginal model.

10.7 For Table 5.2 on admissions decisions for Florida graduate school applicants, let $y_{ig} = 1$ denote a subject in department i of gender g (1 = females, 0 = males) being admitted. The model $\text{logit}[P(Y_{ig} = 1)] = u_i + \alpha + \beta g$ has $\hat{\beta} = 0.163$ $(SE = 0.111)$. The model of form (10.4) that allows the gender effect to vary randomly by department has $\hat{\beta} = 0.176$ $(SE = 0.132)$, with $\hat{\sigma}_v = 0.20$. Interpret and explain why SE for $\hat{\beta}$ is larger for the second model.

10.8 Analyze Table 9.4 with age and maternal smoking as predictors using (**a**) a logistic-normal model, (**b**) a marginal model, (**c**) a transitional model. Summarize your analyses in a short report. Explain how the interpretation of the maternal smoking effect differs for the three models.

10.9 A crossover study compares two drugs on a binary response variable. The study classifies subjects by age as under 60 or over 60. In a GLMM, these two age groups have the same effect comparing the drugs, but the older group has a much larger variance component for its random effects. For the corresponding marginal model, explain why the drug effect for the older group will be smaller than that for the younger group.

10.10 Table 10.11 reports results of a study of fish hatching under three environments. Eggs from seven clutches were randomly assigned to three treatments, and the response was whether an egg hatched by day 10. The three treatments were (1) carbon dioxide and oxygen removed, (2) carbon dioxide only removed, and (3) neither removed. Let π_{it} denote the probability of hatching for an egg from clutch i in treatment t. Model these data using random effects for the clutches. Interpret results.

10.11 For Table 8.13 on premarital and extramarital sex, a cumulative logit model with a random intercept has $\hat{\beta} = 4.134$ with $SE = 0.330$ and $\hat{\sigma} = 2.076$ for the random

Table 10.11 Data on fish hatching for Exercise 10.10.

	Treatment 1		Treatment 2		Treatment 3	
Clutch	No. Hatched	Total	No. Hatched	Total	No. Hatched	Total
1	0	6	3	6	0	6
2	0	13	0	13	0	13
3	0	10	8	10	6	9
4	0	16	10	16	9	16
5	0	32	25	28	23	30
6	0	7	7	7	5	7
7	0	21	10	20	4	20

Source: Thanks to Rebecca Hale, University of North Carolina, Asheville, for these data.

effects. The corresponding cumulative logit marginal model has $\hat{\beta} = 2.51$. Interpret $\hat{\beta}$ for each model. Why are their values so different?

10.12 Refer to the cereal diet and cholesterol study of Table 6.10. Analyze these data with an ordinal GLMM. Summarize results in a short report, showing edited software output as an appendix.

10.13 For the smoking prevention study analyzed in Section 10.4.2, fit the multilevel model that allows interaction between SC and TV. Interpret effects. Compare results to those for the ordinary logistic model that ignores the clustering of observations.

10.14 Refer to Table 9.1 on opinions about legalized abortion in three situations. Using the `Abortion2` data file at the text website that shows the three responses for each subject on a single line, fit the latent class model with $q = 2$, ignoring gender, and show results. Note from the deviance or fitted values that the model is saturated.

10.15 Fit a logistic-normal random intercept model to the carcinoma ratings of Table 10.7 in the `Carcinoma` data file. (*Note*: The data file must be re-constructed to have 7 lines of observations for each slide and indicator columns for the 7 pathologists.) Compare results to those for latent class models in Section 10.5.2.

CHAPTER 11

CLASSIFICATION AND SMOOTHING *

The analyses in this text have mainly utilized generalized linear models, through ML fitting. This chapter presents alternative ways of analyzing categorical data, (1) by using an algorithm that need not specify a functional form for a relationship or (2) by using a predictor in a model that is more complex than a linear predictor or (3) by fitting GLMs using generalizations of ML that can better handle nonstandard situations such as more parameters than observations.

We first introduce methods for *classification* of observations into categories. *Linear discriminant analysis* uses explanatory variables to predict classification on a binary response variable. It yields a linear predictor formula that is similar to that from logistic regression but is more efficient when the explanatory variables have a normal distribution. A method for constructing a *graphical tree* for making such predictions provides a quite different partitioning of the explanatory-variable space that can easily take into account both nonlinearities and interactions. The tree presents a simple representation of the explanatory-variable values that yield each predicted outcome. *Cluster analysis* groups sets of observations on multiple response variables into clusters of like observations.

We then introduce ways of *smoothing* an unknown nonlinear relationship between $E(Y)$ and the explanatory variables. The *generalized additive model* does this by using a predictor that allows curvature of various types in response curves. *Regularization methods* modify ML for ordinary GLMs to give sensible answers in situations that are unstable because of causes such as the number p of explanatory variables being very large. The *lasso* regularization method focuses on identifying the possibly small subset of the explanatory variables

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

that are truly relevant and can yield $\hat{\beta}_j = 0$ for effects that have only weak evidence of existence.

These topics have quite a different nature than others in this book and we attempt only a brief, nontechnical introduction. The end of each section has references for further details.

11.1 CLASSIFICATION: LINEAR DISCRIMINANT ANALYSIS

In Section 4.6.1, we used logistic regression to classify binary observations. We predicted that $y = 1$ whenever the explanatory variables values are such that $\hat{P}(y = 1)$ exceeds some cutpoint, such as 0.50. This section and the following one present alternative ways to predict y by partitioning the set of explanatory variable values into two sets: In one set, the algorithm predicts $y = 1$, which we denote by $\hat{y} = 1$; in the other set, it predicts $y = 0$, which we denote by $\hat{y} = 0$.

The best known such method, called *linear discriminant analysis*, is a simple alternative to logistic regression that also makes predictions using a linear predictor. Recall that ML for logistic regression makes no assumption about the distribution of the explanatory variables and has a likelihood function based merely on assuming a binomial distribution for Y at each set of explanatory-variable values. If actually the explanatory variables have a joint normal distribution, then the linear discriminant analysis under that assumption is a more efficient ML method for making predictions.

11.1.1 Classification with Fisher's Linear Discriminant Function

We denote the set of explanatory variables by $\boldsymbol{x}$, using the notation $\boldsymbol{X}$ when we treat them as random variables having some joint probability distribution. The eminent statistician R.A. Fisher derived the linear combination of $\boldsymbol{x}$ such that its values when $y = 1$ were separated as much as possible from its values when $y = 0$, relative to the variability of the linear-combination values within each y category. Assuming a common covariance matrix for $\boldsymbol{X}$ within each category for y, his method yields a prediction rule based on a linear predictor called *Fisher's linear discriminant function*. In fact, the observations having $\hat{y} = 1$ are equivalently those for which the ordinary least squares regression[1] of an indicator variable for y on $\boldsymbol{x}$ gives a sufficiently high predicted value $\hat{\mu}$ for $E(Y)$.

The boundary for the prediction rule values having $\hat{y} = 1$ and those values having $\hat{y} = 0$ depends on a prior probability π_0 specified for $P(Y = 1)$. For example, suppose a single explanatory variable x has sample means $\bar{x}_0$ when $y = 0$ and $\bar{x}_1$ when $y = 1$. Then, with $\pi_0 = 0.50$, if $\bar{x}_1 > \bar{x}_0$, the prediction $\hat{y} = 1$ if $x > (\bar{x}_1 + \bar{x}_0)/2$, that is, if x is closer to $\bar{x}_1$ than to $\bar{x}_0$, and the prediction is $\hat{y} = 0$ if $x < (\bar{x}_1 + \bar{x}_0)/2$.

Section 4.6.1 used *classification tables* to summarize predictions using a fitted logistic regression model. This type of table can describe the quality of predictions with any method of classification. The true misclassification probabilities tend to be underestimated by predicting observations using the equation to which those observations contributed. Using leave-one-out cross-validation yields less biased estimates; with it, the classification for a particular observation uses the linear discriminant function obtained with the other $n - 1$ observations.

[1] This is the least squares fit for the linear probability model.

11.1.2 Example: Horseshoe Crab Satellites Revisited

In Section 4.6.1, we illustrated classification tables for logistic regression using the horseshoe crab data set, for the model using $x =$ shell width and $c =$ color as predictors of whether a female crab has at least one male satellite. To illustrate linear discriminant analysis, we will use the quantitative darkness scoring (1, 2, 3, 4) for the color levels. Color is discrete rather than continuous, so optimality results of the method for normally distributed explanatory variables do not apply, but we can still use the method to construct a prediction rule. With quantitative scoring of color, violations of a common covariance matrix for $\boldsymbol{X}$ within each category for y may be less severe than when treating color as qualitative with indicator variables.

The R output shown next reports Fisher's linear discriminant function as $0.429x - 0.552c$:

```
-------------------------------------------------------------------------
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
+                     header=TRUE)
> library(MASS)
> lda(y ~ width + color, data=Crabs)
Coefficients of linear discriminants:
           LD1
width  0.42904
color -0.55171
> fit.lda <- lda(y ~ width + color, prior=c(0.5, 0.5), CV=TRUE, data=Crabs)
  # if prior not specified, uses sample proportions in the two categories
> xtabs(~Crabs$y + fit.lda$class) # using cross-validation (CV=TRUE)
       fit.lda$class
Crabs$y  0  1
      0 43 19
      1 39 72
-------------------------------------------------------------------------
```

The predicted $\hat{y} = 1$ when this linear predictor takes a sufficiently large value. With prior value $\pi_0 = 0.50$, one can show that $\hat{y} = 1$ when the least squares fit of the linear probability model, which is $\hat{\pi} = -1.236 + 0.081w - 0.104c$, yields $\hat{\pi} > 0.614$. The output also shows the classification table following leave-one-out cross-validation, for predictions based on this prior value.

Table 11.1 shows the classification table as well as the classification table for logistic regression, using the same cutoff for $\hat{P}(Y = 1)$. The overall proportion of correct

Table 11.1 Classification tables for predictions using discriminant analysis and logistic regression for horseshoe crab data.

	Discriminant Analysis		Logistic Regression		
Actual	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	Total
$y = 1$	72	39	77	34	111
$y = 0$	19	43	21	41	62

predictions was $(72 + 43)/173 = 0.665$ for discriminant analysis and $(77 + 41)/173 = 0.682$ for logistic regression.

11.1.3 Discriminant Analysis Versus Logistic Regression

The linear discriminant function is derived based on assuming a common covariance matrix for $\boldsymbol{X}$ within each category for y. It does not require the assumption that $\boldsymbol{X}$ has a normal distribution, but when that is the case, linear discriminant analysis is optimal for classifying observations. It is then more efficient than logistic regression, potentially considerably more as the groups become more widely separated. This is because it utilizes information about the joint distribution of $\boldsymbol{X}$, which logistic regression ignores.

Often, however, explanatory variables can be far from normally distributed, such as when at least one explanatory variable is qualitative. Also, extreme outliers on $\boldsymbol{x}$ can have a large effect on discriminant analysis, as in ordinary linear regression, but little impact on logistic regression. Therefore, logistic regression is more robust and has a broader scope, as it makes no assumption about a distribution for $\boldsymbol{X}$ and merely assumes a binomial distribution for Y at each value of $\boldsymbol{x}$. Also, logistic regression has the advantage over discriminant analysis of providing direct ways of summarizing effects of explanatory variables, through odds ratios.

Linear discriminant analysis generalizes to multicategory classification, beyond the scope of this book. For further details about linear discriminant analyses, see James et al. (2013, Chapter 4) and Tutz (2011).

11.2 CLASSIFICATION: TREE-BASED PREDICTION

In recent years non-model-based methods have been further developed for predicting response variables using a set of explanatory variables. These methods are often referred to with the terms *statistical learning* or *machine learning*. Rather than relying on a model to summarize effects of explanatory variables on the response variable, such methods are algorithm-driven. Using various criteria, they provide a way of "learning" from the available information on all the variables to estimate the unknown relationship between $E(Y)$ and the explanatory variables $\boldsymbol{x}$. This results in an algorithm for making future predictions of y based solely on values of $\boldsymbol{x}$. The effectiveness of the algorithm is evaluated by its error rate for new samples.

11.2.1 Classification Trees

The *classification tree* method for binary responses provides a graphical tree that depicts the decision process leading to response predictions. This method formalizes a decision process that uses a sequential set of questions about the $\boldsymbol{x}$ values to induce a partitioning of the values that yield each classification prediction for y. Compared with logistic regression and linear discriminant analysis, this classification method is much less restrictive in the form for the predictor decision boundary. The classification tree summarizes binary splits on variables at various stages to determine the prediction. The set of $\boldsymbol{x}$ values for which $\hat{y} = 1$ consists of a set of rectangular regions that can be easily summarized without a linear predictor or other formula.

It is simplest to explain the classification tree method in the context of an example.

11.2.2 Example: A Classification Tree for Horseshoe Crab Mating

We continue from Section 11.1.2 the analysis of the horseshoe crab data using width and quantitative color explanatory variables to predict whether a female crab has satellites. Figure 11.1 shows one possible classification tree. The points on the classification tree at which binary splits occur are called *nodes*. The initial node at the top of the figure containing all the observations is the *root node*. The numbers listed at this root node indicate that 62 crabs had $y = 0$ and 111 crabs had $y = 1$. The nodes beyond which no further splits occur are called *terminal nodes*. Figure 11.1 has 9 nodes, of which 5 are terminal nodes. At each nonterminal node, the left branch has the crabs that satisfy the inequality (e.g., width <25.85 at the root node) and the right branch has the crabs that do not satisfy it.

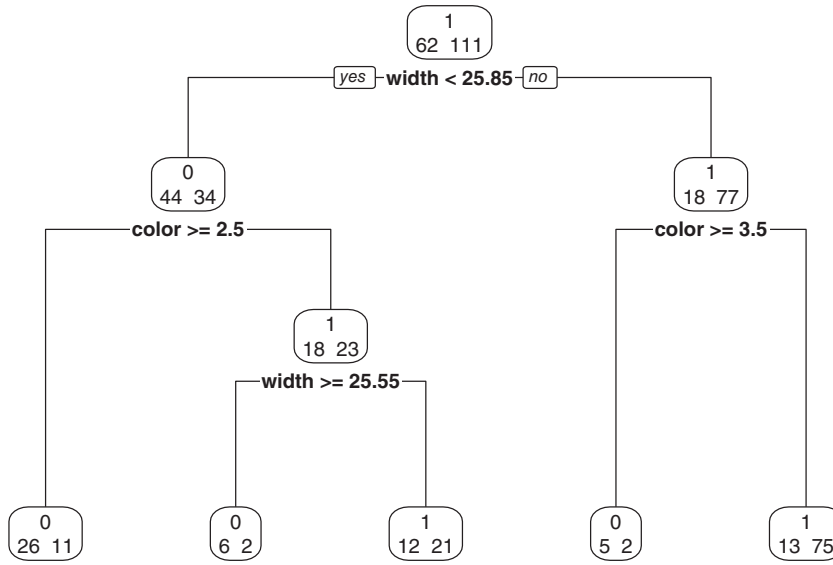


Figure 11.1 Classification tree for horseshoe crab mating. Each node shows the number of crabs with $y = 0$ (no satellites) followed by the number of crabs with $y = 1$. The terminal nodes at the bottom show whether $\hat{y} = 1$ or $\hat{y} = 0$ at that node by the number at the top of the box.

The terminal nodes partition the entire sample into disjoint subsets. Each terminal node has a prediction of 0 or 1 for $\hat{y}$ at the top of its box, for crabs having that a combination of width and color values. These nodes indicate that the horseshoe crabs predicted to have satellites were those in two rectangular regions of the explanatory-variable space that plots width against color: (1) crabs of width $\geq$25.85 cm with color in level 3 or less (medium dark or lighter) and (2) crabs of width $<$25.55 with color in level 2 or less. No crabs were predicted to have satellites that were dark-colored (level 4) or that had width in the interval $25.55 \leq$ width <25.85.

The classification tree summarizes responses to four questions with binary outcomes, with the *yes* response going to the left branch. The questions, together with the counts having each response at a particular node, are:

Q1: Is the width $<$25.85 cm? (78 yes, 95 no)

Q2: Of those having width $\geq$25.85, is color $\geq$3.5 (i.e., in level 4?) (7 yes, 88 no)

Q3: Of those having width $<$25.85, is color in levels (3, 4)? (37 yes, 41 no)

Q4: Of those having width $<$25.85 and color in levels (1, 2), is width $\geq$25.55? (8 yes, 33 no)

For example, one terminal node of the classification tree shows that the 88 crabs with width $\geq$25.85 cm and color in levels (1, 2, 3) were predicted to have satellites; in fact, 75 had satellites but 13 did not. A classification table indicates that this classification tree correctly predicts (without cross-validation) the proportion $96/(96 + 15) = 0.86$ of the crabs that actually had satellites and $37/(37 + 25) = 0.60$ of those that did not.

11.2.3 How Does the Classification Tree Grow?

The method for constructing a binary classification tree uses a *recursive partitioning* algorithm for determining (1) how to choose the splitting variable at each node, (2) how to split a node on the chosen variable, and (3) how to declare a node to be terminal. Without going into detail, we now outline the main ideas. First, binary splits are used instead of multiway splits so the data do not get too fragmented too quickly. In any case, multiway splits can result from a series of binary splits, such as Figure 11.1 does with width.

The classification tree method begins at the root node with the entire sample and first selects the best binary predictor of the response variable. In Figure 11.1, width is split into $<$25.85 and $\geq$25.85 cm. This produces two new nodes, each of which is a candidate for further binary splitting. For an ordinal variable or a quantitative variable such as width, the split takes the form of values falling above versus below a particular level. For a nominal variable, the split is based on ordering the categories by the sample proportions falling in the response category of interest and then using the same criterion to select a cutpoint to separate them into two sets of categories. To find the first binary split, the algorithm forms a classification table for each possible binary split for each explanatory variable. The chosen split satisfies some optimality criterion, such as maximizing the difference between the binomial-likelihood-based deviance for the model with a common probability for all observations and the model allowing two disjoint regions of $\boldsymbol{x}$ values, each having a common probability. The same procedure is then used with each new node.

11.2.4 Pruning a Tree and Checking Prediction Accuracy

The tree can continue growing until there are as many nodes as distinct sets of values of the explanatory variables. In practice, this is overfitting. For a classification tree to perform better for future prediction and not be overfitted, some branches of the tree produced by the basic algorithm can be eliminated. This process is called *pruning*. Ideally, a tree is relatively simple and has good predictive accuracy.

The choice of a *complexity parameter* λ determines the extent of pruning. With $\lambda = 0$ we get the most complex possible tree, whereas as λ increases, more pruning occurs and the tree gets simpler. The choice for λ reflects the fundamental statistical tradeoff between bias and variance that Section 5.1.6 discussed. Fitting the data well (small λ and many terminal nodes) has low bias, whereas having a parsimonious tree (large λ and relatively few terminal nodes) has low variance.

With very small complexity parameter λ, the data are often overfitted. This may be the case for the tree that Figure 11.1 shows, which used $\lambda = 0.02$. Logistic regression and linear discriminant analysis suggested that the probability of having satellites increases as width increases and as color is less dark. For example, a linear discriminant function predicts $\hat{y} = 1$ for relatively large values of $0.429x - 0.552c$. By contrast, the classification tree predicts that for crabs having width $<$25.85 cm and colors (1, 2), those with width $<$25.55 cm

have satellites but those with width >25.55 cm do not. This prediction for the 8 non-dark-colored crabs with width between 25.55 cm and 25.85 cm seems rather anomolous. A tree that is more in accordance with the previous analyses results with greater λ, which prunes to a simpler tree. For instance, $\lambda = 0.07$ yields the simple tree with three terminal nodes that predicts satellites for all crabs of width $\geq$25.85 cm and for all crabs of colors (1, 2) that have width <25.85 cm.

We could select λ to attempt to minimize the cross-validated error rate. However, at each λ, the estimated error rate is itself a random variable. Overfitting is less likely with a *one-standard-error rule*, in which the chosen λ has mean error about one standard error above the minimum. Following is R code for building a classification tree for the horseshoe crab data:

```
------------------------------------------------------------------------------
> Crabs <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat",
+                     header=TRUE)
> library(rpart) # or can use tree package described by James et al. (2013)
> fit <- rpart(y ~ color + width, method="class", data=Crabs)
>              # method="class" for categorical y
> plotcp(fit) # plots error rate by cp = complexity parameter for pruning
> # select leftmost cp with mean error below horizontal line (1SE above min.)
> p.fit <- prune(fit, cp=0.056) # prune with particular value for cp
> library(rpart.plot)
> rpart.plot(p.fit, extra=1, digits=4, box.palette=0) # plots the pruned tree
------------------------------------------------------------------------------
```

11.2.5 Classification Trees Versus Logistic Regression and Discriminant Analysis

Classification trees provide a simple mechanism for using answers to a set of binary explanatory questions to predict a binary response variable. A person can view the tree and clearly see which subjects have $\hat{y} = 1$. Compared with logistic regression and linear discriminant analysis, tree-based classification has the advantage of being easily understandable and useable by practitioners who have little understanding of statistical methods. Also, the trees do not require assumptions about the functional relationship between the response variable and the explanatory variables. In particular, it is easier to detect potentially important interaction structure.

A disadvantage of a classification tree compared with logistic regression modeling and linear discriminant analysis is the lack of smoothness, due to the region of explanatory variable values having $\hat{y} = 1$ being a set of rectangular regions. If a simple linear structure truly describes how the explanatory variables affect the response, as in a logistic regression model having only main effects, the tree will not help us discover this structure. Logistic regression also has the advantage over classification trees (and discriminant analysis) of providing direct ways of summarizing effects of explanatory variables, through odds ratios. Moreover, those effects are conditional on the other explanatory variables, whereas with the classification tree the displayed effects are mixed; the first split refers to a marginal effect, the second to a conditional effect given the first split, and so forth.

Finally, the classification tree method can have low bias but high variance. Classification trees produced by different random samples from a common population can be very different, partly because of its hierarchical nature. Two samples that have a different initial split may end up with very different trees because of the influence of the initial split on the way the tree evolves.[2] Because of this variability, the classification tree method can require rather large n to work effectively. Because of its lack of smoothness and atheoretic nature, many researchers use this method mainly in an exploratory manner. Results of a classification tree, combined with existing theory, can suggest logistic models to use in future research.

In summary, Sections 11.1 and 11.2 suggest that (1) if it seems reasonable to assume that $\boldsymbol{X}$ is approximately normally distributed with common covariance when $y = 0$ and when $y = 1$, simple linear discriminant analysis is sensible for classification; (2) if $\boldsymbol{X}$ may be far from normal but logistic regression seems reasonable, we can use it for classification; (3) if $\boldsymbol{X}$ may interact in unknown ways to determine y but simple rectangular regions are desired for classification, then classification trees are appropriate. For more details about the process of forming classification trees, such as ways of choosing the complexity parameter for the pruning and the role of cross-validation, see James et al. (2013, Chapter 8) and Tutz (2011, Chapter 11).

11.3 CLUSTER ANALYSIS FOR CATEGORICAL RESPONSES

Linear discriminant analysis and classification trees are like logistic regression in using values on explanatory variables to classify observations into well-defined groups that are the categories of the response variable. In some applications, such groups are not identified, but it is still relevant to sort observations into *clusters* of like observations. For example, in "market basket data," a person's observation is a long list of binary indicators in which a particular component indicates whether the person purchased a particular item. A marketing study might want to identify groups of customers with similar buying behavior. Some companies that sell products over the Internet recommend products to people based on a cluster *affinity analysis* that takes into account their purchase history and the history of other people who have bought the same items.

In this section, we present cluster analysis methods for a data file that consists of n observations of p binary response variables. The data file is a $n \times p$ table of indicator variables, where row i shows the p binary responses for observation i, with each response being an indicator that takes value 0 or 1. The goal is to group those n observations into a set of clusters. Those clusters can be regarded as categories of an unknown variable. The number of clusters itself may be unknown.

11.3.1 Measuring Dissimilarity Between Observations

To group together similar observations, clustering methods use a measure of *dissimilarity* between pairs of observations. Ideally, pairs of observations within a cluster have low dissimilarity whereas pairs of observations in different clusters have high dissimilarity. A

[2] Generalizations of classification trees aim to reduce the high variability, such as by averaging many trees (e.g., with *bagging* and *random forests*).

clustering method is characterized by its dissimilarity measure and the algorithm for implementing the clustering.

For observations on p binary variables, Table 11.2 summarizes the similarity and dissimilarity for a particular pair of observations. Of the p variables, a of them take the value 1 for both observations and d of them take the value 0 for both. A simple similarity measure for that pair of observations is the proportion of the p variables that have a match, which is $(a+d)/(a+b+c+d)$. The corresponding dissimilarity index is the proportion for which the outcome differs, which is $(b+c)/(a+b+c+d)$.

Table 11.2 Cross-classification of two observations on p binary response variables, where $p=(a+b+c+d)$.

	Observation 2	
Observation 1	1	0
1	a	b
0	c	d

In some applications, a common response of 1 is more relevant than a common response of 0. With market basket data, for example, each person's observation consists of a very high proportion of 0 entries (i.e., items *not* bought), so necessarily a high proportion of variables has a common outcome. Then, an asymmetric similarity measure may be more relevant. A popular similarity measure of this type is $a/(a+b+c)$, the number of variables coded as 1 for both observations divided by the number of variables that are coded as 1 for either or both observations. The corresponding dissimilarity index is $(b+c)/(a+b+c)$.

11.3.2 Hierarchical Clustering Algorithm and Dendrograms

For a particular dissimilarity measure, one algorithm for performing the clustering creates a *hierarchical* merging of the observations. The clusters at a particular level of the hierarchy result from merging clusters at the next level. At one extreme each observation forms its own cluster and at the other extreme a single cluster contains all the observations. The entire hierarchy portrays an ordered sequence of clusters. We start with each observation as its own cluster and successively merge them. With this *agglomerative clustering*, a step of the algorithm combines into a single cluster the pair of clusters having the smallest average dissimilarity. A tree called a *dendrogram* portrays the process of merging the clusters, as a function of a metric such as the average dissimilarity between clusters being merged.

Any algorithm requires some termination condition for determining the number of clusters k. For example, agglomerative hierarchical clustering can keep combining clusters as long as the average dissimilarity between a pair of clusters to be combined is less than some particular value. An informal way to choose k looks for a natural break point in the dendrogram where the average dissimilarity changes substantially.

With very high-dimensional data, such as market basket data, the challenges to clustering are many, regardless of the type of data. Many irrelevant variables may have the impact of masking clusters, as clusters might exist only for a very small subset of the variables. Dissimilarity measures then are less meaningful, as observations may be close for the most relevant

variables but the curse of dimensionality may put them far apart in high-dimensional space. The clusters found need not reflect a true categorical classification.

11.3.3 Example: Clustering States on Presidential Elections

Table 11.3 shows, for some US states, the political party (Democratic or Republican) that won the electoral votes for that state for each Presidential election between 1980 and 2016. The `Elections` data file at the text website shows results for all states and for the District of Columbia, with indicator variables for the winner (0 = Republican, 1 = Democratic), as illustrated in the `R` code later in this section.

Table 11.3 Statewide data on political party (D = Democratic, R = Republican) winning electoral votes in presidential elections (complete data at text website).

State	1980	1984	1988	1992	1996	2000	2004	2008	2012	2016
Arizona	R	R	R	R	D	R	R	R	R	R
California	R	R	R	D	D	D	D	D	D	D
Colorado	R	R	R	D	R	R	R	D	D	D
Florida	R	R	R	R	D	R	R	D	D	R
Illinois	R	R	R	D	D	D	D	D	D	D
Massachusetts	R	R	D	D	D	D	D	D	D	D
Minnesota	D	R	D	D	D	D	D	D	D	D
Missouri	R	R	R	D	D	R	R	R	R	R
New Mexico	R	R	R	D	D	D	R	D	D	D
New York	R	R	D	D	D	D	D	D	D	D
Ohio	R	R	R	D	D	R	R	D	D	R
Texas	R	R	R	R	R	R	R	R	R	R
Virginia	R	R	R	R	R	R	R	D	D	D
Wyoming	R	R	R	R	R	R	R	R	R	R

For a pair of states, we measure dissimilarity as the number of election outcomes on which the states differ. For example, Arizona and California agree in 4 of the 10 elections, so their dissimilarity is 6. States with identical sets of responses, such as Massachusetts and New York, have dissimilarity values of 0.

The agglomerative algorithm starts with 51 clusters, one for each state and DC. At the first step, states are combined that have the minimum dissimilarity, which in this case consists of states such as Massachusetts and New York that have dissimilarity of 0. At that step, eight clusters of states have dissimilarity of 0 for all pairs within each cluster. At the next step, clusters are combined with the next smallest dissimilarity, such as those two states with California, for which the dissimilarity is 1. By the stage at which only two clusters remain, one cluster has the states (including DC.) that have tended to vote Democratic and the other cluster has the states that have tended to vote Republican.

To more easily portray the cluster-forming process as well as its dendrogram, we redo the analysis using only the data shown in Table 11.3 for 14 states. Figure 11.2 shows the dendrogram. The bottom nodes of the figure are the initial 14 clusters. The top of the dendrogram joins all states into a single cluster. The two-cluster solution, below it, shows the Republican-leaning cluster (Texas, Wyoming, Arizona, Missouri, Colorado, Virginia, Florida, Ohio) and the Democratic-leaning cluster (Minnesota, New Mexico, California,

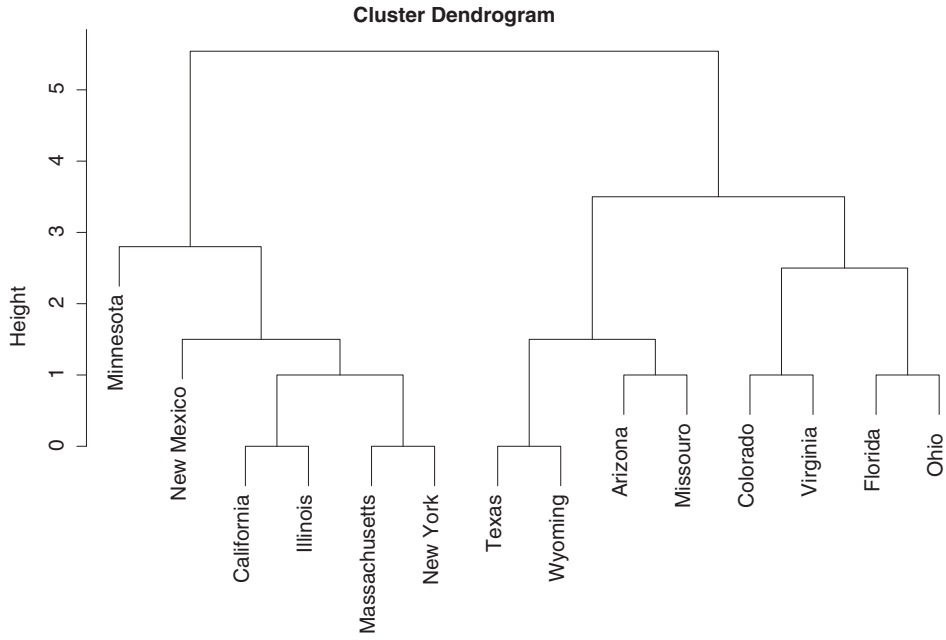


Figure 11.2 Dendrogram for cluster analysis of 14 states according to presidential election results, for data in Table 11.3. The vertical axis (height) measures average dissimilarity.

Illinois, Massachusetts, New York). At the two-cluster step, when Minnesota is joined with five other states, the average dissimilarity between Minnesota and the other five states is $(4 + 3 + 3 + 2 + 2)/5 = 2.8$. Here is R output for this method:

```
-------------------------------------------------------------------------------
> Elections <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                         Elections2.dat", header=TRUE)
> Elections # 0 = Republican, 1 = Democratic winner of Electoral college
            state e1 e2 e3 e4 e5 e6 e7 e8 e9 e10 # 10 election results
1         Arizona  0  0  0  0  1  0  0  0  0   0
2      California  0  0  0  1  1  1  1  1  1   1
3        Colorado  0  0  0  1  0  0  0  1  1   1
4         Florida  0  0  0  0  1  0  0  1  1   0
5        Illinois  0  0  0  1  1  1  1  1  1   1
6   Massachusetts  0  0  1  1  1  1  1  1  1   1
7       Minnesota  1  1  1  1  1  1  1  1  1   1
8        Missouri  0  0  0  1  1  0  0  0  0   0
9       NewMexico  0  0  0  1  1  1  0  1  1   1
10        NewYork  0  0  1  1  1  1  1  1  1   1
11           Ohio  0  0  0  1  1  0  0  1  1   0
12          Texas  0  0  0  0  0  0  0  0  0   0
13       Virginia  0  0  0  0  0  0  0  1  1   1
14        Wyoming  0  0  0  0  0  0  0  0  0   0
> distances <- dist(Elections[, 3:12], method = "manhattan")
```

```
>  # manhattan measures dissimilarity by no. of election outcomes that differ
> democlust <- hclust(distances, "average") # hierarchical clustering
> plot(democlust, labels=Elections$state)
> library(gplots) # heatmap portrays observations with the dendrogram
> heatmap.2(as.matrix(Elections[, 3:12]), labRow = Elections$state,
>   dendrogram="row", Colv=FALSE) # portrays observations with the dendrogram
-------------------------------------------------------------------------------
```

A *heatmap* can portray all the observations together with the dendrogram. See Figure 11.3.

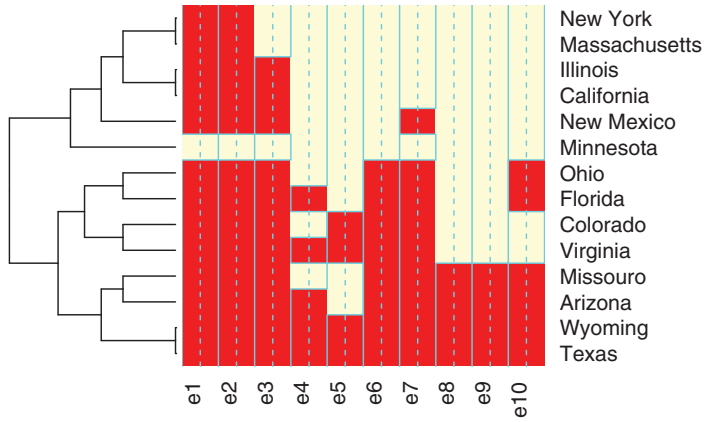


Figure 11.3 Heatmap showing observations (shaded = Republican victory) and dendrogram for Table 11.3.

See Bartholomew et al. (2008, Chapter 2) and James et al. (2013, Section 10.3) for further details about cluster analysis methods.

11.4 SMOOTHING: GENERALIZED ADDITIVE MODELS

This section presents a model that has a predictor that is more general than a linear predictor, replacing the linear effect of an explanatory variable by a smooth function of that variable. In basing analyses on a more general structure, we have less potential for incorrect conclusions because of model misspecification. However, in some ways the demands are greater: we need to choose among a potentially infinite number of smooth forms relating the response variable to the explanatory variables; the number of parameters is also then potentially much larger and overfitting is a danger.

11.4.1 Generalized Additive Models

The GLM generalizes the ordinary linear model to permit non-normal distributions and link functions of the mean. A further generalization replaces the linear predictor by additive smooth functions of the explanatory variables. With it, the GLM structure for observation i and its mean $\mu_i = E(Y_i)$ modeled as a function of p explanatory variables generalizes to

$$g(\mu_i) = s_1(x_{i1}) + s_2(x_{i2}) + \cdots + s_p(x_{ip}),$$

where s_j is an unspecified smooth function[3] of explanatory variable j. Like GLMs, this model specifies a link function g and a probability distribution for Y. The resulting model is called a *generalized additive model*, symbolized by GAM.

The GLM is the special case in which each s_j is a linear function. Also possible is a mixture of explanatory terms of various types, with some s_j as smooth functions, others as linear functions such as in GLMs, and others as indicator variables to include qualitative factors. The GAM fit may suggest that a GLM is adequate, or it may suggest ways to improve on an ordinary linear predictor.

GAMs have the advantage over GLMs of greater flexibility. Using them, we may discover patterns we would miss with ordinary GLMs, and we obtain potentially better estimates of mean responses. A disadvantage of GAMs and other smoothing methods, compared with GLMs, is the loss of simple interpretability for describing the effect of an explanatory variable or obtaining confidence intervals for those effects. Therefore, it is more difficult to judge when an effect has substantive importance. Also, because any smoothing method has potentially a large number of parameters, it can require a very large n to estimate the functional form accurately.

Even if you plan mainly to use GLMs, GAMs are helpful for exploratory analysis. For instance, for binary responses, scatterplots are not very informative. Plotting the fitted smooth function for a predictor may reveal a general trend without assuming a particular functional relation.

11.4.2 Example: GAMs for Horseshoe Crab Data

For the horseshoe crab data, Figure 3.3 in Section 3.3.3 showed the trend relating a female crab's number of male satellites to her shell width. The smooth curve shown is the fit of a GAM, assuming a Poisson distribution and using the log link function, obtained using the `R` code in Section 3.3.3.

For the binary response of whether a female horseshoe crab has at least one satellite, Figure 4.2 in Section 4.1.3 plotted these data against width. That figure also showed a curve based on smoothing the data using a GAM, assuming a binomial response and logit link function, with the `R` code in Section 4.1.3. This curve shows a roughly increasing trend and is more informative than viewing the binary data alone.

11.4.3 How Much Smoothing? The Bias/Variance Tradeoff

Smoothing methods such as the GAM require input from the data analyst to control the degree of smoothness imposed on the data. As usual, this choice relates to the statistical tradeoff between bias and variance. Greater smoothness has the effect of decreasing the variance in estimating characteristics of interest (such as probabilities), but at the cost of increasing bias.

The degree of smoothness for each s_j in the additive predictor of a GAM results from selecting an *effective df* for the term. For instance, a smooth function having $df = 3$ is similar in overall complexity to a third-degree polynomial, which has three parameters (besides the intercept). Choosing a df value determines how smooth the resulting GAM fit looks.

[3] A common way to smooth uses *splines*. The cubic spline has separate cubic polynomials over sets of disjoint intervals, joined together smoothly at boundaries of those intervals.

Larger df values result in less smooth functions (more "wiggling"). It is usually sensible to try various degrees of smoothing to find one that smooths the data appropriately. With insufficient smoothing, the data are overfitted with a highly wiggly function. However, with too much smoothing, informative patterns are suppressed.

With a chosen effective df value for each s_j in the additive predictor, we can conduct inference about those terms. Any model fit has a residual deviance, which reflects the assumed distribution for Y. As in GLMs, we can compare deviances for nested models to test whether a model gives a significantly better fit than a simpler model. When we let the GAM procedure determine an optimal effective df rather than select it ourselves a priori, however, we should regard statistical significance in an informal manner.

11.4.4 Example: Smoothing to Portray Probability of Kyphosis

A study[4] investigated risk factors for kyphosis, which is severe forward flexion of the spine following corrective spinal surgery. We analyze the binary response y for kyphosis ($1 =$ present, $0 =$ absent) in terms of $x =$ the age (in months) at the time of operation, for $n = 40$ children. Figure 11.4 shows the data, which are in the data file `Kyphosis` at the text website. The figure suggests that at the very low and very high age levels, most observations have kyphosis absent, but we should be cautious in making conclusions about trends with such a small n.

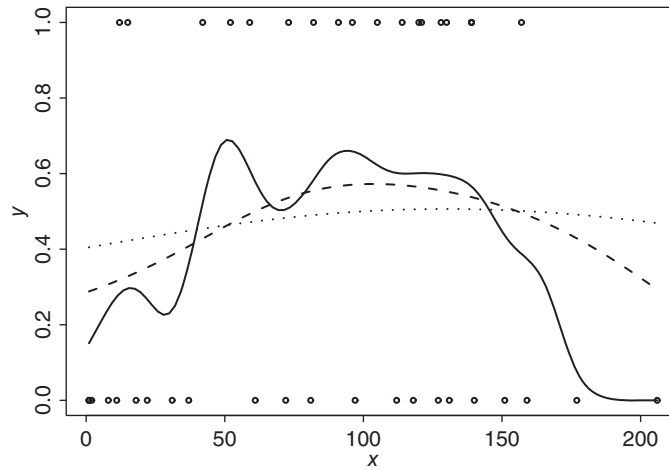


Figure 11.4 Data and smoothed estimates of probability of kyphosis as a function of $x =$ age (in months).

Many non-model-based methods are available for smoothing data, such as *kernel smoothing* and *loess*, and Figure 11.4 also shows three results from using a kernel smoothing method. Depending on the degree of smoothness imposed, the prediction for $P(Y = 1)$ can be quite smooth (nearly parabolic) or it can be more irregular than justified with such small n. Let us see what kind of smoothing a GAM suggests. For fitting a GAM, we treat the data

[4] Described on p. 282 of Hastie and Tibshirani (1990).

as binomial with a logit link. We next fit[5] models that have successively linear, quadratic, and cubic complexity for the smooth function, as shown in the following R output:

```
------------------------------------------------------------------------------
> Kyphosis <- read.table("http://www.stat.ufl.edu/~aa/cat/data/
+                        Kyphosis.dat", header=TRUE)
> Kyphosis # Data file for n=40 at text website
     x y
1   12 1
...
40 206 0
> library(gam)
> gam.fit1 <- gam(y ~ s(x, df=1), family=binomial, data=Kyphosis)
> gam.fit2 <- gam(y ~ s(x, df=2), family=binomial, data=Kyphosis)
> gam.fit3 <- gam(y ~ s(x, df=3), family=binomial, data=Kyphosis)
> anova(gam.fit1, gam.fit2, gam.fit3) # fitting models sequentially
Analysis of Deviance Table
Model 1: y ~ s(x, df=1) # linear complexity
Model 2: y ~ s(x, df=2) # quadratic
Model 3: y ~ s(x, df=3) # cubic
  Resid.  Df  Resid. Dev       Df  Deviance  Pr(>Chi)
1         38      54.504
2         37      49.216   0.9999    5.2880    0.0215
3         36      48.231   1.0002    0.9852    0.3210
> plot(y ~ x, xlab="Age", ylab="Presence of Kyphosis", data=Kyphosis)
> curve(predict(gam.fit2, data.frame(x=x), type="resp"), add=TRUE)
------------------------------------------------------------------------------
```

Comparison of deviances suggests that quadratic fits better than linear, but cubic is not significantly better than quadratic. A plot of the GAM fitted values with quadratic complexity (requested in this output) suggests using a logistic model with a quadratic term. That model fit is similar graphically and in the residual deviance. You can check that adding a quadratic age term to the logistic regression model using age as the predictor provides an improved fit (decrease in deviance giving P-value $= 0.012$).

For further details about generalized additive models, see Hastie and Tibshirani (1990), Hastie et al. (2009, Chapter 9), James et al. (2013, Chapter 7), Tutz (2011, Chapter 10), Wood (2017), and Yee (2015, Chapter 4).

11.5 REGULARIZATION FOR HIGH-DIMENSIONAL CATEGORICAL DATA (LARGE *p*)

High-dimensional data are increasingly common in applications in various disciplines, such as genomics, biomedical imaging, market basket data, and portfolio allocation in finance.

[5] The VGAM and mgcv libraries in R can also fit GAMs.

In genomics, for example, such applications include classifying tumors by using microarray gene expression data, predicting a clinical prognosis by using gene expression data, and detecting differential expression (change between two or more conditions) in many thousands of genes. High-dimensional data are not well handled by the traditional ML model-fitting methods presented in this book. We now discuss issues in modeling when the number of explanatory variables p is very large, sometimes even with $p > n$. Certain issues are vital yet difficult, such as how to select explanatory variables from an enormous set when nearly all of them are expected to have either no effect or a practically insignificant effect.

When we use ordinary GLMs with high-dimensional data, the ML estimates can be highly unstable or not even exist. *Regularization methods* are ways of modifying the ML estimates to give sensible answers in such situations.

11.5.1 Penalized-Likelihood Methods and L_q-Norm Smoothing

The penalized likelihood method adds a term to the log-likelihood function such that the resulting estimates that maximize it are smoothings of the ordinary ML estimates. For a model with log-likelihood function $L(\boldsymbol{\beta})$ in terms of a set of parameters $\boldsymbol{\beta}$, we maximize

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) - s(\boldsymbol{\beta}),$$

where s is a function such that $s(\boldsymbol{\beta})$ decreases as elements of $\boldsymbol{\beta}$ are smoother in some sense. The penalized likelihood estimate typically shrinks the ML estimates toward 0. Among its positive features are a reduction in prediction error and sensible values when the ML estimate may not exist or is infinite or is badly affected by multicollinearity.

Section 5.4 introduced a version of penalized likelihood for logistic regression that is useful for reducing bias. We used it for the problematic situation of complete separation, which results in infinite parameter estimates. Another penalized-likelihood approach that applies to any GLM uses the L_q*-norm* smoothing function

$$s(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|^q$$

for some power $q \geq 0$ and a smoothing parameter $\lambda \geq 0$. Two cases for q are especially popular. L_2*-norm* penalized likelihood ($q = 2$), in which the penalty is proportional to the sum of squared parameter values, is called *ridge regression* in the context of ordinary linear models. L_1*-norm* penalized likelihood ($q = 1$), in which the penalty is proportional to the sum of absolute parameter values, is called the *lasso*.

For any q, λ is a smoothing parameter that reflects the bias/variance tradeoff. Ordinary ML estimates result from $\lambda = 0$. Increasing λ results in greater shrinkage of $\{\hat{\beta}_j\}$ toward 0, potentially reducing the variance but increasing the bias. In implementing this method, we express the model in terms of standardized variables. Then, the smoothing function treats each variable in the same way, and the degree of smoothing does not depend on the choice of scaling.

11.5.2 Implementing the Lasso

The lasso, which is the penalized likelihood using $s(\boldsymbol{\beta}) = \lambda \sum_j |\beta_j|$, is an acronym for *least absolute shrinkage and selection operator*. Equivalently, the lasso method maximizes the likelihood function subject to the constraint that $\sum_j |\hat{\beta}_j| \leq \lambda^*$ for a constant λ^* that is inversely related to the smoothing parameter λ. As λ increases or as λ^* decreases, this method shrinks more of the $\hat{\beta}_j$ completely to zero.

A plot of the lasso estimates as a function of λ summarizes how $\{\hat{\beta}_j\}$ reach 0 and corresponding explanatory variables drop out of the linear predictor as λ increases. To choose λ using cross-validation, for each λ value in a grid, we fit the model to part of the data and then check the goodness of the predictions for y in the remaining data. With k-fold cross-validation, we do this k times (for k typically about 10), each time leaving out the fraction $1/k$ of the data and predicting those y values using the model fit from the rest of the data. The summary measure of prediction error is the sample mean prediction error for the k runs, for a measure of prediction error. We could then select the λ that has the lowest sample mean prediction error. However, at each λ, the sample mean prediction error is itself a random variable. An alternative choice uses the *one-standard-error rule*, in which the chosen λ has a mean prediction error that is one standard error above the minimum, in the direction of greater regularization. Such a choice is less likely to overfit the model. Whichever value we select for λ, we then apply that value with the lasso method for all the data. For a particular data set, different data analysts will typically report different lasso results, as the chosen λ is a random variable that depends on the data-splitting for the cross-validation.

A disadvantage of the lasso approach is that it can overly penalize β_j that are truly large. Also, the lasso estimators $\{\hat{\beta}_j\}$ do not have approximate normal sampling distributions and can be highly biased, making statistical inference difficult. Software for the lasso does not report standard errors, which are potentially misleading with such highly biased estimates.[6] Which of L_1 norm and L_2 regularization performs better in terms of bias and variance for estimating the true $\{\beta_j\}$ depends on their values. When p is large but only a few $\{\beta_j\}$ are practically different from 0, L_1 norm tends to perform better, because many $\{\hat{\beta}_j\}$ may equal 0. When $\{\beta_j\}$ do not vary dramatically in substantive size, L_2 norm regularization tends to perform better.

11.5.3 Example: Predicting Opinion on Abortion with Student Survey

The `Students` data file at the text website shows responses to a questionnaire by 60 social science graduate students in an introductory Statistics course at the University of Florida. The variables are summarized in Exercise 5.4 in Chapter 5. Here we use the 14 binary and quantitative variables to predict a subject's response about whether abortion should be legal in the first three months of pregnancy (1 = yes, 0 = no).

When we use all 14 explanatory variables as main effects in a logistic regression model, the results reflect the relatively large p with n of only 60. The following `R` output shows

[6] The distance of an estimator from a parameter depends on both variance and bias; the *mean squared error* equals the variance plus the squared bias. Measures of variability are available with a Bayesian implementation of the lasso (e.g., with `R` package `monomvn`).

that although the likelihood-ratio test has strong evidence against H_0: $\beta_1 = \cdots = \beta_{14} = 0$, only two explanatory variables are significant at the 0.05 level in Wald tests, and the same is true with likelihood-ratio tests:

```
-------------------------------------------------------------------------------
> Students <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Students.dat",
+                        header=TRUE)
> fit <- glm(abor ~ gender + age + hsgpa + cogpa + dhome + dres + tv + sport +
+  news + aids + veg + ideol + relig + affirm, family=binomial, data=Students)
> summary(fit)
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  10.10142    10.89140    0.927    0.3537
gender        1.00216     1.86553    0.537    0.5911
age          -0.07834     0.12748   -0.615    0.5389
hsgpa        -3.73445     2.80932   -1.329    0.1837
cogpa         2.51127     3.73991    0.671    0.5019
dhome         0.00056     0.00068    0.821    0.4116
dres         -0.33882     0.29538   -1.147    0.2514
tv            0.26598     0.25316    1.051    0.2934
sport         0.02721     0.25515    0.107    0.9151
news          1.38688     0.69868    1.985    0.0471 # LR P-value = 0.0003
aids          0.39668     0.56637    0.700    0.4837
veg           4.32135     3.86146    1.119    0.2631
ideol        -1.63779     0.78925   -2.075    0.0380 # LR P-value = 0.0010
relig        -0.72457     0.78207   -0.926    0.3542
affirm       -2.74815     2.68988   -1.022    0.3069
---
    Null deviance: 62.719  on 59  degrees of freedom
Residual deviance: 21.368  on 45  degrees of freedom
AIC: 51.368
> 1 - pchisq(62.719 - 21.368, 59 - 45) # LR test that all 14 betas = 0
[1] 0.0001566051
-------------------------------------------------------------------------------
```

Four explanatory variables (news, ideol, relig, affirm) show significance when used as the sole predictor. We could use the methods discussed in Section 5.1 to build a model. As an alternative, let us see what happens when we use the lasso, implemented in R with the `glmnet` package, which operates on the standardized variables. Here, we show the coefficients obtained with the smoothing parameter value of $\lambda = 0.1268$ suggested with the one-standard-error rule in a particular lasso fit with cross-validation.

```
-------------------------------------------------------------------------------
> attach(Students)
> x <- cbind(gender, age, hsgpa, cogpa, dhome, dres, tv, sport, news, aids,
>            veg, ideol, relig, affirm) # explanatory var's for lasso
> library(glmnet)
> fit.lasso <- glmnet(x, abor, alpha=1, family="binomial") # alpha=1 is lasso
> plot(fit.lasso, "lambda")
> set.seed(1) # a random seed to implement cross-validation
```

```
> cv.glmnet(x, abor, alpha=1, family="binomial", type.measure="class")
  $lambda.min # best lambda by 10-fold cross-validation
  [1] 0.06610251 # this is a random variable, and changes from run to run
  $lambda.1se # lambda suggested by one-standard-error rule, also a r.v.
  [1] 0.1267787
> coef(glmnet(x, abor, alpha=1, family="binomial", lambda=0.1267787))
                  s0
(Intercept)  2.36711 # all 12 lasso estimates that are not shown equal 0
ideol       -0.25994
relig       -0.18311
-------------------------------------------------------------------------------
```

That fit has only *ideol* (political ideology, scaled from 1 = very liberal to 7 = very conservative) and *relig* (how often you attend religious services, with 0 = never, 1 = occasionally, 2 = most weeks, 3 = every week) as explanatory variables, with estimated effects -0.260 and -0.183.

The value $\lambda = 0.0661$ gave the minimum cross-validated mean squared error. Its fit adds *news* as a predictor, but with a coefficient of only 0.140, much less than the ML estimate of 1.387 for the full model. For this lasso fit with $\lambda = 0.0661$, the estimated *ideol* and *relig* effects are -0.424 and -0.360, not shrunk quite as much towards 0 from the ML estimates as they are by the lasso fit with $\lambda = 0.1268$ based on the one-standard-error rule.

Figure 11.5 shows how the lasso estimates change as λ increases on the log scale. The values at the left end of the plot, with the smallest shown value for $\log(\lambda)$, are very close to the ML estimates (for which $\lambda = 0$). The *ideol* estimate shrinks toward 0 from the ML value of -1.638, becoming 0 when $\log(\lambda) \geq -1.5$ (i.e., $\lambda \geq 0.222$). The ML estimate of -0.725 for *relig* decreases to 0 when $\log(\lambda) \geq -1.73$.

We next compare the lasso-fit estimates with $\lambda = 0.0661$ to the ML estimates for the model with *ideol*, *relig*, and *news* as explanatory variables. The ML estimates are

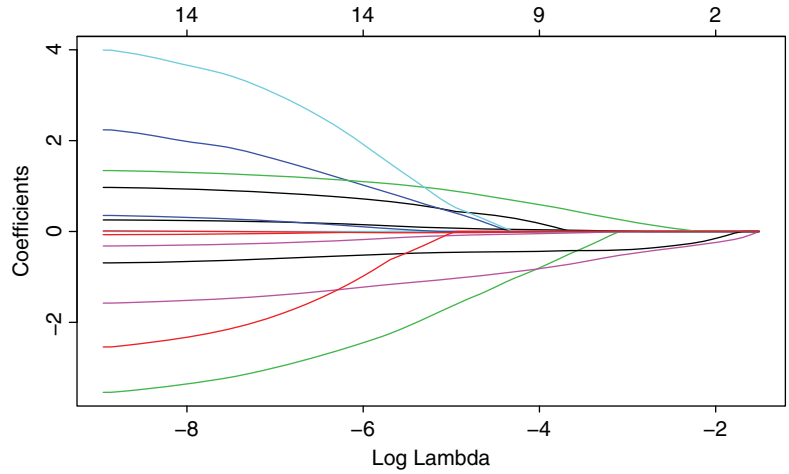


Figure 11.5 Plot of lasso model parameter estimates for predicting opinion on legalized abortion using student survey data, as a function of smoothing parameter $\log(\lambda)$.

substantially larger in absolute value than the lasso estimates, although *relig* does not have a statistically significant effect, adjusting for the other two variables.

```
-------------------------------------------------------------------------------
> summary(glm(abor ~ ideol + relig + news, family = binomial, data=Students))
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   3.5205      1.2513    2.814   0.00490
ideol        -1.2515      0.4671   -2.679   0.00738 # lasso est. = -0.424
relig        -0.7198      0.4982   -1.445   0.14854 # lasso est. = -0.360
news          1.1292      0.4574    2.469   0.01356 # lasso est. =  0.140
---
     Null deviance: 62.719  on 59  degrees of freedom
Residual deviance: 29.791  on 56  degrees of freedom
AIC: 37.791
-------------------------------------------------------------------------------
```

11.5.4 Why Shrink ML Estimates Toward 0?

To methodologists who commonly use estimators that are unbiased or approximately so, methods such as penalized likelihood that shrink $\{\hat{\beta}_j\}$ toward 0 can seem counterintuitive. Why is it that such shrinkage can be effective? In studies having a large number of explanatory variables, often most of them have no effect or very minor effects. An example is genetic association studies, which simultaneously consider each of possibly thousands of genes for the association between the genetic expression levels and the response of whether a person has a particular disease. Unless n is very large, by ordinary sampling variability, the ML estimates $\{\hat{\beta}_j\}$ tend to be much larger in absolute value than the true values. This tendency is exacerbated when we consider only statistically significant values.

Shrinkage such as occurs with penalized likelihood methods tends to move the ML estimates closer to the true values. This is yet another example of the bias/variance tradeoff. Introducing a penalty function results in biased estimates but benefits from reducing the variance. For more details about penalized likelihood methods, see James et al. (2013, Chapter 6).

11.5.5 Issues in Variable Selection (Dimension Reduction)

When the number of predictors p is large, selecting a sensible set of explanatory variables for a model is challenging. A fundamental assumption needed for methods to perform well is *sparse structure*, that is, relatively few explanatory variables in the model. Removing variables that have little if any relevance can ease interpretability and decrease prediction errors. With large p, ordinary ML fitting may not even be possible. For a binary response, complete or quasi-complete separation often occurs when p exceeds a particular point, resulting in some infinite estimates. Even when finite estimates exist, they may be imprecise because of multicollinearity. Moreover, using a large number of explanatory variables in a model runs the risk of overfitting the data. Future predictions will then tend to be poorer than those obtained with a more parsimonious model.

Variable selection methods for large p fall roughly into three types. One approach adopts model variable-elimination methods, such as stepwise methods and the lasso. A second

approach replaces the p explanatory variables by a small set of artificial variables that are linear combinations of the original explanatory variables but account for most of their variabillity. A third approach attempts to identify the relevant effects using standard significance tests but with an adjustment for the possibly huge number of tests conducted. Approaches (1) and (3) can reduce dramatically the data dimensionality by eliminating the many explanatory variables that do not have strong evidence of an effect, and in approach (2) the number of newly created variables can also be much less than p.

As in ordinary model selection, the first approach of using an automated variable selection algorithm such as backward elimination has pitfalls, even more so when p is large. For example, for the set of explanatory variables that have no true effect, the maximum sample correlation with the response can be quite large. Also, there can be spurious collinearity among the predictors or spurious correlation between an important predictor and a set of unimportant predictors, because of the high dimensionality. Other criteria exist for identifying an optimal subset of explanatory variables, such as minimizing AIC. With large p, though, it is not feasible to check a high percentage of the possible subsets of predictors, and the danger remains of identifying an effect as important when it is actually not. The lasso also yields zero weight for many explanatory variables in the prediction equation, but in an objective way that does not depend on which variables were previously eliminated.

The second approach that explicitly performs dimension reduction is *principal component analysis*. This method replaces the p explanatory variables by fewer linear combinations of them (the "principal components") that are uncorrelated. The first principal component is the linear combination that has the largest possible variance. Each succeeding component has the largest possible variance under the constraint that it is uncorrelated with the preceding components. A small number of principal components often explains a high percentage of the original variability. The components depend on the scaling of the original variables, so when they measure inherently different characteristics they are standardized before beginning the process. A disadvantage, especially with large p, is that it may be difficult to interpret the principal components. For details, see James et al. (2013, Chapter 6) and Bartholomew et al. (2008, Chapter 5).

The third approach uses standard significance tests to judge the effects of the explanatory variables, but with an adjustment for multiplicity. One way to do this uses the *false discovery rate* (FDR).

11.5.6 Controlling the False Discovery Rate

With a large number of statistical inferences when p is large, such as testing H_0: $\beta_j = 0$ at the $\alpha = 0.05$ level for each explanatory variable, the danger exists of making many Type I errors when few variables truly have an effect. It is common then to use a method that instead sets α to be an overall error rate of at least one incorrect inference. When the number of planned significance tests equals t, the *Bonferroni* method ensures that α is an upper bound for the overall Type I error rate, by using α/t as the size for each test; that is, H_0 is rejected in a particular test when the P-value $\leq \alpha/t$. When t is enormous, the power may be very low for establishing significance with any individual inference. It can be difficult to discover effects that truly exist, especially if those effects are weak, but, in the absence of an adjustment, most significant results found could be Type I errors, especially when the number of true non-null effects is small.

A multiple inference method that addresses this issue controls the *false discovery rate* (FDR). In the context of significance testing, this is the expected proportion of the rejected null hypotheses ("discoveries") that are erroneously rejected (i.e., that are actually true – "false discoveries"). A simple algorithm can ensure FDR $\leq \alpha$ when applied with t independent tests.[7] Let $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(t)}$ denote the ordered P-values for the t tests. The method rejects hypotheses (1), ..., (j^*), where j^* is the maximum j for which $P_{(j)} \leq j\alpha/t$. The most significant test compares $P_{(1)}$ to α/t and has the same decision as in the ordinary Bonferroni method, but then the other tests have less conservative requirements. When some hypotheses are false, the FDR method tends to reject more of them than the Bonferroni method, which focuses solely on controlling the family-wise error rate. The actual FDR for this method is bounded above by α times the proportion of rejected hypotheses that are actually true. This bound is α when the null hypothesis is always true.

The statisticians (Yoav Benjamini and Yosef Hochberg) who introduced this method in 1995 illustrated the FDR for a study about myocardial infarction. For the $t = 15$ hypotheses tested in the study, the ordered P-values ($P_{(1)}, P_{(2)}, \ldots, P_{(15)}$) were

$$0.0001, 0.0004, 0.0019, 0.0095, 0.020, 0.028, 0.030,$$
$$0.034, 0.046, 0.32, 0.43, 0.57, 0.65, 0.76, 1.00.$$

With $\alpha = 0.05$, these are compared with $j\alpha/t = j(0.05)/15 = (0.0033)j$, starting with $j = 15$. The maximum j for which $P_{(j)} \leq (0.0033)j$ is $j = 4$, for which $P_{(4)} = 0.0095 < (0.0033)4$. Therefore, the hypotheses with the four smallest P-values are rejected. By contrast, the Bonferroni approach with family-wise error rate 0.05 compares each P-value to $0.05/15 = 0.0033$ and rejects only three of these hypotheses. The Bonferroni method yields an adjusted P-value that is 15 times the ordinary one (with a maximum of 1.0), whereas the FDR method has smaller adjusted P-values after the first:

```
-------------------------------------------------------------------------
> pvals <- c(0.0001, 0.0004, 0.0019, 0.0095, 0.020, 0.028, 0.030,0.034,
+            0.046, 0.32, 0.43, 0.57, 0.65, 0.76, 1.00)
> p.adjust(pvals, method=c("bonferroni"))
 [1] 0.0015 0.0060 0.0285 0.1425 0.3000 0.4200 0.4500 0.5100 0.6900
[10] 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000
> p.adjust(pvals, method=c("fdr"))
 [1] 0.0015 0.0030 0.0095 0.0356 0.0600 0.0638 0.0638 0.06380.07667
[10] 0.4800 0.5864 0.7125 0.7500 0.8143 1.0000
-------------------------------------------------------------------------
```

The FDR method is especially useful in applications in which a very small proportion of the variables are expected to truly have an effect. Because of its lessened conservatism and improved power compared with family-wise inference methods such as the Bonferroni, controlling FDR is a sensible strategy to employ in exploratory research involving large-scale significance testing. A place remains for traditional family-wise inference methods in follow-up validation studies involving the smaller numbers of effects found to be significant in the exploratory studies.

[7] The method also works with tests that are positively dependent in a certain sense.

11.5.7 Large *p* also Makes Bayesian Inference Challenging

Dealing with large p is also challenging for Bayesian inference, perhaps even more so than for frequentist inference. The impact of forming prior distributions for a very large number of parameters may differ from what you intuitively expect. For example, even if you pick a very diffuse prior, the effect may depend strongly on which diffuse prior you choose.

Suppose the response distribution is multinomial with p outcome categories having probabilities $\{\pi_j\}$. Here, for simplicity, we discuss large-p challenges[8] without any reference to explanatory variables. In practice, similar issues arise when the number of multinomial categories is of any size but the number of explanatory variables is large. Let n_j denote the total number of observations in outcome category j, with $n = \sum_j n_j$. The beta distribution that serves as a conjugate prior distribution for a binomial parameter extends to the *Dirichlet distribution* for multinomial parameters. With hyperparameters $\{\alpha_j\}$, the posterior density of $\{\pi_j\}$ is also Dirichlet. The posterior mean of π_j is $(n_j + \alpha_j)/(n + \sum_k \alpha_k)$. The impact of the prior is essentially to add α_j observations to category j for all j before forming a sample proportion for a category. Most applications use a common value α for $\{\alpha_j\}$, so the impact is to smooth in the direction of the equi-probability model. The Dirichlet prior with $\alpha = 1$ corresponds to a uniform prior distribution over the probability simplex. This seems diffuse, but it corresponds to adding p observations and then forming sample proportions. This is considerable when p is large.

For example, suppose we observe which in a list of $p = 1000$ books is selected as the favorite book by each of $n = 100$ people. With uniform prior, the posterior mean of π_j is $(n_j + 1)/(n + p) = (n_j + 1)/1100$. When book j receives 1 of the 100 observations, the posterior mean estimate for that book is $2/1100 = 0.002$, shrinking the sample proportion of 0.010 for the book toward the equi-probability value of 0.001. This seems like a reasonable estimate. However, what if instead book j receives all 100 observations? The posterior mean estimate is then $101/1100 = 0.092$. This shrinks much more from the sample proportion value of 1.0 than we are likely to believe is sensible. Even though the prior distribution is quite diffuse, it has a strong impact on the results.

Much more diffuse priors can be more effective, yielding posterior results for each parameter that are more similar to what we would obtain if p were small. In applications for which we believe that many effects are truly 0, one Bayesian approach uses a *spike-and-slab* prior distribution that mixes the ordinary prior distribution with a discrete one that places all its probability at 0.

For further details about regularization methods, see Hastie et al. (2009, Chapter 3) and James et al. (2013, Chapter 6).

EXERCISES

11.1 Refer to the use of linear discriminant analysis by R.A. Fisher in 1936 for Iris flower data, as discussed in the article "Iris flower data set" at Wikipedia. Conduct a linear discriminant analysis using the data given there for the versicolor and virginica species (which are in the Iris data file at the text website), with sepal length and petal length as explanatory variables. Report the linear discriminant function. Show the leave-one-out cross-validated classification table for $\pi_0 = 0.50$.

[8] Here, the actual number of multinomial parameters is $p - 1$ becasuse $\sum_j \pi_j = 1$.

11.2 For the `Crabs` data file with $y =$ whether a female horseshoe crab has at least one satellite, conduct a linear discriminant analysis. Use all four explanatory variables, treating color and spine condition as quantitative. Report the linear discriminant function and compare classification tables based on leave-one-out cross-validation for this method and for logistic regression with the same explanatory variables.

11.3 Consider the horseshoe crab classification tree in Figure 11.1.

a. Sketch a plot of width against color. Based on the information at the terminal nodes, show the rectangular regions for which $\hat{y} = 1$. Compare to the regions for the simpler classification tree mentioned in Section 11.2.4 that uses $\lambda = 0.07$.

b. Constructing your own classification trees with these data for various λ values, investigate how the pruned tree depends on λ. Show the tree that seems most sensible to you.

11.4 Refer to the previous exercise. Use the classification tree method with all four explanatory variables, treating color and spine condition as quantitative. Explain how you pruned the tree, interpret it, and explain how (if at all) it conflicts with the results from the logistic model-building of Section 5.1.5.

11.5 A classification tree predicted whether, over a one-year period, elderly subjects participating in an assisted-living program disenroll from the program to enter a nursing home.[9] The sample consisted of 4654 individuals who had been enrolled in the program for at least a year and who did not die during that one-year period. If available online at your library, read the article cited in the footnote.

a. Specifiy the four questions asked in the classification tree.

b. In this tree, the authors treated the two types of misclassifications as having different costs — 13 times as high to predicting that someone would remain in the program who actually left it than to predicting that someone would leave the program who actually stayed. At a terminal node, the prediction taken is the response category that has the lowest misclassification cost. For the node of 931 subjects of age $>$83, of whom 112 disenrolled and 819 stayed, show that the misclassification cost is lower if we predict that they all disenroll; this is therefore the prediction for this terminal node.

c. Form a classification table from results in the classification tree. Show that the tree correctly predicts the proportion 0.40 of those who actually disenrolled and 0.79 of those who remained.

11.6 Refer to the Presidential elections cluster analysis in Section 11.3.3.

a. The observations were the same for New Jersey and Pennsylvania as llinois, and the same for North Carolina as Virginia. Conduct a cluster analysis, showing the dendrogram, using these three states together with the 14 states in Table 11.3. Interpret.

b. Repeat the cluster analysis of Section 11.3.3 without the Presidential elections of 1980 and 1984, in which nearly all states favored the Republican candidate (Ronald Reagan). Show the dendrogram and compare results to the analysis that used those elections.

[9] Noe et al., *Chance* **22**, no. 4: 58–62 (2009).

11.7 Conduct a cluster analysis for the complete `Elections` data file at the text website. Show the two-cluster solution and interpret.

11.8 For the horseshoe crab data set, plot the binary response of the satellite presence against the crab's weight. Also plot a curve based on smoothing the data using a generalized additive model, assuming a binomial response and logit link. Interpret.

11.9 For the horseshoe crab data set, plot the satellite count against the crab's weight. Also plot a curve based on smoothing the data using a generalized additive model, assuming a Poisson response and log link. Interpret.

11.10 For the data analyzed in Sections 5.3.4, 5.4.2, and 5.4.4 on risk factors for endometrial cancer, compare the results shown there with those you obtain with the lasso by selecting λ using cross-validation with (**a**) the lowest sample mean prediction error, (**b**) the one-standard-error rule.

11.11 For the `Students` data file, model $y =$ whether you are a vegetarian, with the 13 other explanatory variables used in Section 11.5.3. For ordinary logistic regression, show that this model fits better than the null model, but no Wald tests are significant. Use the lasso, with λ to minimize the sample mean prediction error in cross-validation and with the one-standard-error rule. Compare estimates to those from ordinary logistic regression.

11.12 For the myocardial infarction example in Section 11.5.6, use the FDR method to test significance, with $\alpha = 0.10$. Compare results to the Bonferroni method.

11.13 Using the *spam* data set at the website `web.stanford.edu/~hastie/ElemStatLearn` for Hastie et al. (2009), use two methods presented in this chapter to select explanatory variables to classify whether a given email is spam. Explain how you implemented the methods, form classification tables, and summarize results in a short report, with edited software output in an appendix.

11.14 Project: Go to a site with large data files, such as the UCI Machine Learning Repository (`archive.ics.uci.edu/ml`) or Yahoo! Webscope (`webscope.sandbox.yahoo.com`). Find a data set of interest to you that has a categorical response variable. Use at least one method presented in this chapter to analyze the data. Summarize your analyses in a short report.

CHAPTER 12

A HISTORICAL TOUR OF CATEGORICAL DATA ANALYSIS *

We conclude by providing a historical overview of the evolution of methods for categorical data analysis (CDA). The beginnings of CDA were often shrouded in controversy. Key figures in the development of statistical science made groundbreaking contributions, but these statisticians were often in heated disagreement with one another.

The Pearson–Yule Association Controversy

Much of the early development of methods for CDA took place in England. It is fitting that we begin our historical tour in London in 1900, because in that year Karl Pearson (1857–1936) introduced his chi-squared statistic (X^2). Pearson's motivation for developing the chi-squared test included testing whether outcomes on a roulette wheel in Monte Carlo varied randomly and testing statistical independence in two-way contingency tables.

Much of the CDA literature in the early 1900s consisted of vocal debates about appropriate ways to summarize association. Pearson's approach assumed that continuous bivariate distributions underlie cross-classification tables. He argued that one should describe association by approximating a measure, such as the correlation, for the underlying continuum. In 1904, Pearson introduced the term *contingency* as a "measure of the total deviation of the classification from independent probability," and he introduced measures to describe its extent and to estimate the correlation.

George Udny Yule (1871–1951), an English contemporary of Pearson's, took an alternative approach in his study of association between 1900 and 1912. He believed that many

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

categorical variables are inherently discrete. Yule defined measures, such as the odds ratio, directly using cell counts without assuming an underlying continuum. Discussing one of Pearson's measures that assumes underlying normality, Yule stated "at best the normal coefficient can only be said to give us in cases like these a hypothetical correlation between supposititious variables. The introduction of needless and unverifiable hypotheses does not appear to me a desirable proceeding in scientific work." Yule also showed the potential discrepancy between marginal and conditional associations in contingency tables, later studied by E.H. Simpson in 1951 and now called *Simpson's paradox*.

Karl Pearson did not take kindly to criticism, and he reacted negatively to Yule's ideas. For example, Pearson claimed that the values of Yule's measures were unstable, since different collapsings of $r \times c$ tables to 2×2 tables could produce quite different values. In 1913, Pearson and D. Heron filled more than 150 pages of *Biometrika*, a journal he co-founded and edited, with a scathing reply to Yule's criticism. In a passage critical also of Yule's well-received book *An Introduction to the Theory of Statistics*, they stated

> If Mr. Yule's views are accepted, irreparable damage will be done to the growth of modern statistical theory. . . . [His measure] has never been and never will be used in any work done under his [Pearson's] supervision. . . . We regret having to draw attention to the manner in which Mr. Yule has gone astray at every stage in his treatment of association, but criticism of his methods has been thrust on us not only by Mr. Yule's recent attack, but also by the unthinking praise which has been bestowed on a text-book which at many points can only lead statistical students hopelessly astray.

Pearson and Heron attacked Yule's "half-baked notions" and "specious reasoning" and concluded that Yule would have to withdraw his ideas "if he wishes to maintain any reputation as a statistician."

Half a century after the Pearson–Yule controversy, Leo Goodman and William Kruskal of the University of Chicago surveyed the development of measures of association for contingency tables and made many contributions of their own. Their 1979 book, *Measures of Association for Cross Classifications*, reprinted their four influential articles on this topic. Initial development of many measures occurred in the 1800s, such as the use of the relative risk by the Belgian social statistician Adolphe Quetelet in 1849. The following quote from an article by M.H. Doolittle in 1887 illustrates the lack of precision in early attempts to quantify *association* even in 2×2 tables.

> Having given the number of instances respectively in which things are both thus and so, in which they are thus but not so, in which they are so but not thus, and in which they are neither thus nor so, it is required to eliminate the general quantitative relativity inhering in the mere thingness of the things, and to determine the special quantitative relativity subsisting between the thusness and the soness of the things.

R.A. Fisher's Contributions

Karl Pearson's disagreements with Yule were minor compared to his later ones with Ronald A. Fisher (1890–1962). Using a geometric representation, in 1922 Fisher introduced *degrees of freedom* to characterize the family of chi-squared distributions. Fisher claimed that for tests of independence in $r \times c$ tables, X^2 had $df = (r-1)(c-1)$. By contrast, in 1900 Pearson had argued that, for any application of his statistic, df equalled the number of cells

minus 1, or $rc - 1$ for two-way tables. Fisher pointed out, however, that estimating hypothesized cell probabilities using estimated row and column probabilities resulted in additional $(r-1)+(c-1)$ constraints on the fitted values, thus affecting the distribution of X^2.

Not surprisingly, Pearson reacted critically to Fisher's argument. He stated "I hold that such a view (Fisher's) is entirely erroneous, and that the writer has done no service to the science of statistics by giving it broad-cast circulation in the pages of the *Journal of the Royal Statistical Society*. . . . I trust my critic will pardon me for comparing him with Don Quixote tilting at the windmill; he must either destroy himself, or the whole theory of probable errors, for they are invariably based on using sample values for those of the sampled population unknown to us." Pearson claimed that using row and column sample proportions to estimate unknown probabilities had negligible effect on large-sample distributions. Fisher was unable to get his rebuttal published by the Royal Statistical Society and he ultimately resigned his membership.

Statisticians soon realized that Fisher was correct. For example, in an article in 1926, Fisher provided empirical evidence to support his claim. Using 11,688 2×2 tables randomly generated by Karl Pearson's son, E.S. Pearson, he found a sample mean of X^2 for these tables of 1.00001; this is much closer to the 1.0 predicted by his formula for $E(X^2)$ of $df = (r-1)(c-1) = 1$ than Pearson's $rc - 1 = 3$. Fisher maintained much bitterness over Pearson's reaction to his work. In a later volume of his collected works, writing about Pearson, he stated "If peevish intolerance of free opinion in others is a sign of senility, it is one which he had developed at an early age."

Fisher also made good use of CDA methods in his applied work. For example, he was also a famed geneticist. In one article, Fisher used Pearson's goodness-of-fit test to test Mendel's theories of natural inheritance. Calculating a summary P-value from the results of several of Mendel's experiments, he obtained an unusually large value ($P = 0.99996$) for the right-tail probability of the reference chi-squared distribution. In other words X^2 was so small that the fit seemed *too* good, leading Fisher in 1936 to comment "the general level of agreement between Mendel's expectations and his reported results shows that it is closer than would be expected in the best of several thousand repetitions. . . . I have no doubt that Mendel was deceived by a gardening assistant, who knew only too well what his principal expected from each trial made." In a letter written at the time, he stated "Now, when data have been faked, I know very well how generally people underestimate the frequency of wide chance deviations, so that the tendency is always to make them agree too well with expectations."

In 1934 the fifth edition of Fisher's classic text *Statistical Methods for Research Workers* introduced "Fisher's exact test" for 2×2 contingency tables. In his 1935 book *The Design of Experiments*, Fisher described the tea-tasting experiment (Section 2.6.2) based on his experience at an afternoon tea break while employed at Rothamsted Experiment Station. Other CDA-related work of his included proposing linear discriminant analysis and showing how to assign scores to rows and columns of a contingency table to maximize the correlation.

Logistic Regression

Model building for categorical responses did not receive much attention before the 1970s, with a few exceptions. For instance, in 1935 Chester Bliss popularized the probit model for applications in toxicology dealing with a binary response, and Fisher showed how to find ML estimates for it using the iteratively reweighted least squares algorithm that is today

called *Fisher scoring*. In 1944, the physician and statistician Joseph Berkson introduced the term "logit" for the transformation $\log[\pi/(1-\pi)]$ of a binomial parameter and showed that the logistic regression model fitted similarly to the probit model.

In the early 1970s, work by the Danish statistician and mathematician Georg Rasch sparked an enormous literature on item response models. The most important of these is the logit model with subject and item parameters, now called the *Rasch model* (Section 10.2.4). This work was highly influential in the psychometric community of northern Europe (especially in Denmark, the Netherlands, and Germany) and spurred many generalizations in the educational testing community in the United States.

The extension of logistic regression to multicategory responses received occasional attention before 1970, but substantial work after that date. For nominal responses, early work was mainly in the econometrics literature. In 2000, Daniel McFadden won the Prize in Economic Sciences in Memory of Alfred Nobel for his work in the 1970s and 1980s on the discrete-choice model (Section 6.1.6). Cumulative logit models received some attention starting in the 1960s and 1970s, but did not become popular until an article by Peter McCullagh in 1980 provided a Fisher scoring algorithm for ML fitting of a more general model for cumulative probabilities allowing a variety of link functions.

Other major advances with logistic regression dealt with its application to case-control studies in the 1970s and the conditional ML approach to model fitting for those studies and others with numerous nuisance parameters. Biostatisticians Norman Breslow and Ross Prentice at the University of Washington had a strong influence on this. The conditional approach was later exploited in small-sample exact inference in a series of papers by Cyrus Mehta and Nitin Patel, who later founded CyTel Software and developed *StatXact* and *LogXact* software.

Perhaps the most far-reaching contribution was the introduction by British statisticians John Nelder and R.W.M. Wedderburn in 1972 of the concept of *generalized linear models*. This unifies the logistic and probit regression models for binomial responses with loglinear models for Poisson or negative binomial responses and with long-established regression and ANOVA models for normal responses.

More recently, attention has focused on fitting logistic regression models to correlated responses for clustered data. One strand of this is marginal modeling of longitudinal data, proposed by Kung-Yee Liang and Scott Zeger at Johns Hopkins in 1986 using generalized estimating equations (GEE). Another strand is generalized linear mixed models, including multilevel models.

Multiway Contingency Tables and Loglinear Models

Research work on loglinear models and other methods for analyzing association and interaction structure evolved between about 1950 and 1975. At the University of Chicago, Leo Goodman was the most prolific contributor to this advancement of CDA methodology. He and his PhD students Shelby Haberman and Clifford Clogg wrote a series of groundbreaking articles in statistics and social science journals on loglinear and related logit models, specialized models for square tables (e.g., quasi independence) and for ordinal data, and latent class models.

Simultaneously, related research on ML methods for loglinear-logit models occurred at Harvard University by students of Frederick Mosteller (such as Stephen Fienberg) and William Cochran. In the early 1950s, William Cochran had dealt with a variety of

important topics in CDA, such as introducing a generalization (Cochran's Q) of McNemar's test for comparing proportions in several matched samples, partitioning chi-squared statistics, and proposing a test of conditional independence for $2 \times 2 \times K$ tables. Much of the later research was inspired by problems arising in analyzing large, multivariate data sets in the National Halothane Study. That study investigated whether halothane was more likely than other anaesthetics to cause death due to liver damage. A landmark book in 1975 by Yvonne Bishop, Stephen Fienberg, and Paul Holland, *Discrete Multivariate Analysis*, helped to introduce loglinear models to the general statistical community.

Research at the University of North Carolina by Gary Koch and several colleagues was highly influential in the biomedical sciences. Their research developed weighted least squares (WLS) methods for categorical data models. An article in 1969 by Koch with J. Grizzle and F. Starmer popularized this approach. In later articles, Koch and colleagues applied WLS to problems for which ML methods are difficult to implement, such as the analysis of repeated categorical measurement data.

Certain loglinear models with conditional independence structure provide *graphical models* for contingency tables. These relate to the conditional independence graphs that Section 7.4.1 used. An article by John Darroch, Steffen Lauritzen, and Terry Speed in 1980 was the genesis of much of this work.

Final Comments

Methods for CDA continue to be developed. In the past thirty years, an active area of new research has been the modeling of clustered data, such as using GLMMs and marginal models. In particular, multilevel (hierarchical) models have become increasingly popular. Recently, the Bayesian approach has seen renewed interest because of the development of methods for numerically evaluating posterior distributions for increasingly complex models. Another active area of research, largely outside the realm of traditional modeling, is the development of algorithmic methods for huge data sets with large numbers of variables, such as described in the previous chapter.

The above discussion provides only a sketchy overview of the development of CDA. Further details and references for technical articles and books appear in Agresti (2013).

APPENDIX: SOFTWARE FOR CATEGORICAL DATA ANALYSIS

All major statistical software has procedures for categorical data analyses. The text has shown examples with R, and this appendix discusses how to use SAS, Stata, and SPSS. We present simplistic examples without much discussion about available options, but we provide references for further details. See also www.stat.ufl.edu/~aa/cda/cda.html.

For certain analyses, specialized software provides methods not available in the major packages. An example is StatXact (Cytel Software, Cambridge, MA), which provides exact inferential analyses for categorical data. Among its procedures are small-sample confidence intervals for differences and ratios of proportions and for odds ratios, and Fisher's exact test and its generalizations for $r \times c$ and three-way tables. Its companion program LogXact performs exact conditional logistic regression.

A.1 R FOR CATEGORICAL DATA ANALYSIS

R is free software to which users can add functions for implementing new methodology. At www.r-project.org you can download R and find documentation. Helpful resources for a great variety of methods for categorical data include a manual by Laura Thompson at www.stat.ufl.edu/~aa/cda for my more advanced text (Agresti 2013), the books by Bilder and Loughin (2015) and Kateri (2014), and the websites mentioned in Section 1.6. For its use with the methods presented in this book, see the examples in the text discussions.

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

A.2 SAS FOR CATEGORICAL DATA ANALYSIS

In learning SAS, you can get help at sites such as

```
support.sas.com/rnd/app/da/stat/procedures/CategoricalDataAnalysis.html
support.sas.com/documentation/onlinedoc/stat
```

For details about SAS for categorical data analyses, see specialized SAS publications such as Allison (2012) and Stokes et al. (2012).

The main procedures (PROCs) for categorical data analyses are FREQ, GENMOD, LOGISTIC, PROBIT, CATMOD, GLIMMIX, and NLMIXED. PROC FREQ computes chi-squared tests of independence, measures of association and their estimated standard errors, and exact tests of independence in $r \times c$ tables. PROC GENMOD fits generalized linear models, cumulative logit models for ordinal responses, and can perform GEE analyses for marginal models. PROC LOGISTIC provides ML fitting of binary response models, cumulative logit models for ordinal responses, baseline-category logit models for nominal responses, and conditional logistic regression. It incorporates model selection procedures, regression diagnostic options, and exact conditional inference. PROC GLIMMIX and PROC NLMIXED fit generalized linear mixed models.

For convenience, many of the following examples enter data in the form of the contingency table displayed in the text. In practice, data files usually have ungrouped data at the subject level. All SAS statements must end with a semicolon. The data follow the DATALINES statement, one line per subject, unless the INPUT statement ends with @@. Input of a variable as characters rather than numbers requires an accompanying $ label in the INPUT statement.

Chapters 1–2: Introduction and Contingency Tables

With binomial counts of successes and failures, PROC FREQ can provide confidence limits for a binomial proportion. Table A.1 shows code, with 9 successes in 10 trials. Options include the score-test-based confidence interval (WILSON, named after the statistician who first proposed it), Agresti–Coull interval (AC), and the Jeffreys Bayesian posterior interval (JEFFREYS). The keyword BINOMIAL and the EXACT statement yields binomial and large-sample normal tests (e.g., here to test H_0: $\pi = 0.4$).

Table A.1 SAS code for statistical inference for a proportion.

```
-------------------------------------------------------------------------------
data clinical;
   input response $ count;
datalines;
no   1
yes  9
;
proc freq data=clinical; weight count;
   tables response / binomial(wilson ac jeffreys) alpha=.05;
proc freq data=clinical; weight count; exact binomial;
   tables response / binomial (P = 0.4);
-------------------------------------------------------------------------------
```

PROC FREQ creates contingency tables with the TABLES statement, ordering row and column categories alphanumerically. To use instead of the order in which the categories appear in the data set (e.g., to treat the variable properly in an ordinal analysis), use the ORDER=DATA option. The WEIGHT statement is needed when you enter the contingency table with cell counts (i.e., grouped data) instead of subject-level (ungrouped) data. PROC FREQ can conduct chi-squared tests of independence (CHISQ option), show its estimated expected frequencies (EXPECTED), provide a wide assortment of measures of association and their standard errors (MEASURES), and provide ordinal statistic (2.6) for a correlation test (CMH1). Table A.2 uses SAS to analyze Table 2.4. You can also perform chi-squared tests using PROC GENMOD to fit the loglinear model of independence (Section 7.1.1), as also shown in Table A.2. Its RESIDUALS option provides cell residuals. The output labeled *Std Pearson Residual* is the standardized residual (2.5).

Table A.2 SAS code for the chi-squared test, measures of association, and residuals with Political Party ID and gender data from Table 2.4.

```
-------------------------------------------------------------------------------
data table;
   input gender $ party $ count;
datalines;
female dem  495
...
male indep  498
  ;
proc freq  order=data; weight count;
   tables gender*party / chisq expected measures cmh1;
proc genmod  order=data;  class gender party;
   model count = gender party / dist=poi link=log residuals;
-------------------------------------------------------------------------------
```

With PROC FREQ, for 2×2 tables the MEASURES option in the TABLES statement provides confidence intervals for the odds ratio and the relative risk, and the RISKDIFF option provides intervals for the proportions and their difference. Using RISKDIFF(CL=(MN)) gives the interval based on inverting a score test. The EXACT statement can provide various exact analyses. These include Fisher's exact test and its generalization for $r \times c$ tables, treating variables as nominal, with the keyword FISHER. The OR keyword gives the odds ratio and its large-sample and small-sample confidence intervals. Table A.3 analyzes Table 2.7.

Table A.3 SAS code for Fisher's exact test and confidence intervals for odds ratio for tea-tasting data in Table 2.7.

```
-------------------------------------------------------------------------------
data fisher;
   input poured guess count @@;
datalines;
  1 1 3    1 2 1    2 1 1    2 2 3
;
proc freq;  weight count;
   tables poured*guess / measures;
   exact fisher or / alpha=.05;
-------------------------------------------------------------------------------
```

Chapters 3–5: Generalized Linear Models and Logistic Regression

PROC GENMOD fits GLMs using ML or Bayesian methods. It specifies the response distribution in the DIST option (*poi* for Poisson, *bin* for binomial, *mult* for multinomial, *negbin* for negative binomial) and specifies the link in the LINK option. The LRCI option provides profile likelihood confidence intervals. The TYPE3 option provides likelihood-ratio tests for each parameter. Table A.4 illustrates by fitting Poisson and negative binomial loglinear models for Table 3.2.

Table A.4 SAS code for Poisson regression and negative binomial regression for horseshoe crab satellite counts of Table 3.2.

```
-------------------------------------------------------------------------------
data crab;
   input  color  spine  width  satell  weight;
datalines;
 2  3  28.3  8  3.05
...
 2  2  24.5  0  2.00
;
proc genmod;
   model satell = width / dist=poi link=log lrci type3;
proc genmod;
   model satell = width / dist=negbin link=log;
-------------------------------------------------------------------------------
```

Table A.5 uses PROC GENMOD to fit linear probability and logistic models as GLMs for Table 3.1. With grouped data, the response in the model statements takes the form of the number of successes divided by the number of cases.

Table A.5 SAS code for binary GLMs with grouped data for snoring data in Table 3.1.

```
-------------------------------------------------------------------------------
data glm;
   input snoring disease total;
datalines;
  0 24 1379
...
  5 30 254
;
proc genmod; model disease/total = snoring / dist=bin link=identity;
proc genmod; model disease/total = snoring / dist=bin link=logit;
-------------------------------------------------------------------------------
```

PROC LOGISTIC can also fit logistic regression models. With binary data entry, GENMOD and LOGISTIC order the levels alphanumerically, forming the logit with (1, 0) responses as $\log[P(Y = 0)/P(Y = 1)]$. Invoking the procedure with DESCENDING following the PROC name reverses the order. PROC LOGISTIC has a built-in check of whether ML estimates exist by detecting complete separation of data points. The PLCL option provides profile likelihood confidence intervals. The INFLUENCE option provides residuals and diagnostic measures, and options exist for stepwise selection of variables. The CTABLE option gives a classification table, with the cutoff point specified by PPROB. In GENMOD or LOGISTIC, a CLASS statement requests that an explanatory variable be treated as a

qualitative factor by setting up indicator variables. By default, in GENMOD the parameter estimate for the last category of a factor equals 0. In LOGISTIC, estimates sum to zero. That is, indicator variables take the coding $(1, -1)$ of 1 when in the category and -1 when not, for which parameters sum to 0. The option PARAM=REF in the CLASS statement in LOGISTIC requests (1, 0) indicator variables with the last category as the default reference level. Putting REF=FIRST next to a variable name requests its first category as the reference level. Predicted probabilities and lower and upper 95% confidence limits for the probabilities are shown in a plot with the PLOTS=ALL option or with PLOTS=EFFECT. The ROC curve is produced using the PLOTS=ROC option. Table A.6 shows logistic regression analyses for Table 3.2. Following the first LOGISTIC model statement, it requests predicted probabilities and lower and upper 95% confidence limits for the probabilities.

Table A.6 SAS code for logistic regression models for ungrouped horseshoe crab data of Table 3.2.

```
-------------------------------------------------------------------------------
data crab;
   input  color  spine  width  weight  satell  y;

 2  3  28.3  3.05  8  1
...
 2  2  24.5  2.00  0  0
;
proc genmod descending; class color;
   model y = width color / dist=bin link=logit lrci type3 obstats;
proc logistic descending;
   model y = width;
   output out=predict  p=pi_hat lower=LCL  upper=UCL;
proc print  data = predict;
ods graphics on; ods html;
proc logistic descending plots=all; class color spine / param=ref;
   model y = width weight color spine / selection=backward;
proc gam plots=components(clm commonaxes);  * generalized additive model;
   model y (event='1') = spline(width) / dist=binary; * smooth data;
ods html close; ods graphics off;
-------------------------------------------------------------------------------
```

Table A.7 uses GENMOD and LOGISTIC to fit a logistic model with qualitative predictors to the grouped data from Table 4.2. In GENMOD, the OBSTATS option

Table A.7 SAS code for logistic modeling of grouped survey data in Table 4.2.

```
-------------------------------------------------------------------------------
data pot;
   input gender $ race $ y n;
datalines;
 White Female 420 1040
...
 Other Male 32 94
;
proc genmod; class gender race;
   model y/n = gender race / dist=bin type3 lrci residuals obstats;
proc logistic; class gender race / param=ref;
   model y/n = gender race / aggregate scale=none clparm=both clodds=both;
-------------------------------------------------------------------------------
```

provides various *observation statistics*, including predicted values and their confidence limits. The RESIDUALS option requests residuals such as the standardized residuals (labeled *Std Pearson Residual*). In LOGISTIC, the CLPARM=BOTH and CLODDS=BOTH options provide Wald and profile likelihood confidence intervals for parameters and odds ratio effects of explanatory variables. With AGGREGATE SCALE=NONE in the model statement, LOGISTIC reports Pearson and deviance goodness-of-fit tests; it forms groups by aggregating data into the possible combinations of explanatory variable values, without overdispersion adjustments.

Table A.8 uses PROC GENMOD for the Bayesian analysis described in Section 5.4.2. Variances can be set for normal priors (e.g., VAR = 100), the length of the Markov chain can be set by NMC, and DIAGNOSTICS=MCERROR gives the amount of Monte Carlo error in the parameter estimates at the end of the MCMC fitting process. The penalized likelihood approach of Section 5.4.3 for reducing bias in estimation of logistic regression parameters is available with the FIRTH option in PROC LOGISTIC.

Table A.8 SAS code for Bayesian and penalized likelihood modeling of Table 5.4 on endometrial cancer.

```
-------------------------------------------------------------------------------
data endometrial;
   input nv pi eh hg ;
nv2 = nv - 0.5; pi2 = (pi-17.3797)/9.9978; eh2 = (eh-1.6616)/0.6621;
datalines;
  0 13 1.64 0
  0 16 2.26 0
...
  0 33 0.85 1
;
proc genmod descending;
   model hg = nv2 pi2 eh2 /  dist=bin link=logit;
bayes coeffprior=normal (var=100) diagnostics=mcerror nmc=1000000;
proc logistic descending;
   model hg = nv2 pi2 eh2 / firth clparm=pl ;
-------------------------------------------------------------------------------
```

Exact conditional logistic regression is available in PROC LOGISTIC and in PROC GENMOD with the EXACT statement. You can obtain ordinary and mid P-values as well as confidence limits for each model parameter.

Chapters 6–7: Multicategory Logit Models and Loglinear Models

PROC LOGISTIC fits baseline-category logit models using the LINK= GLOGIT option. The final response category is the default baseline for the logits. Table A.9 fits a model to Table 6.1. PROC MDC fits multinomial discrete choice models.

PROC GENMOD can fit the proportional odds version of cumulative logit models using the DIST=MULTINOMIAL and LINK=CLOGIT options (CPROBIT for cumulative probit). Table A.10 fits it to the mental impairment data file introduced in Section 6.3.4. When the number of response categories exceeds two, by default PROC LOGISTIC fits this model. Both procedures have graphic capabilities of displaying the cumulative probabilities as a function of a predictor (with PLOTS = EFFECT), at fixed values of other predictors. With

Table A.9 SAS code for baseline-category logit models with alligator data in Table 6.1.

```
-------------------------------------------------------------------------------
data gator;
   input length choice $ ;
datalines;
  1.24 I
...
  3.89 F
;
proc logistic;
   model choice = length / link=glogit;
-------------------------------------------------------------------------------
```

Table A.10 SAS code for the cumulative logit model with proportional odds structure for mental impairment data file.

```
-------------------------------------------------------------------------------
data mental;
   input impair ses life;
datalines;
 1 1 1
...
 4 0 9
;
proc genmod ;
   model impair = life ses / dist=multinomial link=clogit lrci type3;
proc logistic;
   model impair = life ses;
proc logistic;
   model impair = life ses /  unequalslopes;
-------------------------------------------------------------------------------
```

the UNEQUALSLOPES option, PROC LOGISTIC can fit the model without the proportional odds structure.

You can fit adjacent-categories logit models in SAS by fitting equivalent baseline-category logit models. For an example, see the SAS file at `www.stat.ufl.edu/~aa/cda/cda.html`.

Table A.11 uses GENMOD to fit the loglinear model (AC, AM, CM) to Table 7.1 on alcohol, cigarette, and marijuana use by treating it as a GLM.

Table A.11 SAS code for fitting the loglinear model to substance use data of Table 7.1.

```
-------------------------------------------------------------------------------
data drugs;
   input a c m count;
datalines;
  1 1 1 911
...
  2 2 2 279
;
proc genmod; class a c m;
   model count = a c m a*m a*c c*m / dist=poi link=log lrci type3 obstats;
-------------------------------------------------------------------------------
```

Table A.12 uses GENMOD to fit the linear-by-linear association model to Table 7.10, with column scores (1, 2, 4, 5). The defined variable *assoc* represents the cross-product of row and column scores, which has β parameter as coefficient in the model.

Table A.12 SAS code for fitting the linear-by-linear association model to Table 7.10.

```
-------------------------------------------------------------------------------
data sex;
   input premar birth u v count @@;  assoc = u*v;
datalines;
 1 1 1 1  81   1 2 1 2  68   1 3 1 4 60   1 4 1 5 38
...
;
proc genmod; class premar birth;
   model  count = premar birth assoc / dist=poi link=log;
-------------------------------------------------------------------------------
```

Chapter 8: Matched Pairs

Table A.13 analyzes Table 8.1. The AGREE option in PROC FREQ provides the McNemar chi-squared statistic for binary matched pairs and Cohen's kappa and weighted kappa with SE values. The MCNEM keyword in the EXACT statement provides a small-sample binomial version of McNemar's test. PROC CATMOD can provide the confidence interval for the difference of proportions. The code forms a model for the marginal proportions in the first row and the first column, specifying a matrix in the model statement that has an intercept parameter (the first column) that applies to both proportions and a slope parameter that applies only to the second; hence the second parameter is the difference between the second and first marginal proportions.

Table A.13 SAS code for McNemar's test and comparing proportions for matched samples in Table 8.1.

```
-------------------------------------------------------------------------------
data matched;
   input taxes living count @@;
datalines;
 1 1 227   1 2 132   2 1 107    2 2 678
;
proc freq; weight count;
   tables taxes*living / agree;  exact mcnem;
proc catmod; weight count;
   response marginals;
   model taxes*living =  (1 0 ,
                          1 1 ) ;
-------------------------------------------------------------------------------
```

It is also possible to get this confidence interval with the GEE methods of Chapter 9, using PROC GENMOD with the REPEATED statement as shown in Table A.14 for an ungrouped data file. Replacing the identity link by the logit link yields the estimate for the marginal model having an effect that is the log odds ratio comparing the marginal distributions.

Table A.15 shows how to conduct a Wald test of marginal homogeneity for nominal variables, illustrated using the coffee market share data of Table 8.5.

Table A.14 SAS code estimating difference of correlated binomial proportions for Table 8.1.

```
-------------------------------------------------------------------------------
data matched;
   input person topic y; * topic: 1=taxes, 0=living;  * y: 1=yes, 0=no;
datalines;
1     1  1
1     0  1
...
1144  1  0
1144  0  0
;
proc genmod data=matched; class person;
   model y = topic / dist=binomial link=identity;
   repeated subject=person / type=indep corrw;
-------------------------------------------------------------------------------
```

Table A.15 SAS code showing the test of marginal homogeneity for Table 8.5.

```
-------------------------------------------------------------------------------
data coffee;
   input first $ second $ count ;
datalines;
    highpt highpt 93
    highpt tasters 17
...
    brim brim 27
;
proc catmod; weight count; response marginals;
    model first*second = _response_ /freq;
    repeated time 2;
-------------------------------------------------------------------------------
```

Table A.16 shows a comparison of ordinal margins for Table 8.6, using the GEE methods of Chapter 9 with a cumulative logit model for an ungrouped data file.

Table A.16 SAS code showing a cumulative logit model comparison of margins of Table 8.6.

```
-------------------------------------------------------------------------------
data ordinal;
   input person topic  y;
*topic: 0=Recycle,1=DriveLess; *y: 1=always, 2=often, 3=sometimes, 4=never;
datalines;
 1     0  1
 1     1  1
...
 1230  0  4
 1230  1  4
;
proc genmod data = ordinal; class person;
   model y = topic / dist=multinomial link=cumlogit type3;
   repeated subject=person / type=indep corrw;
-------------------------------------------------------------------------------
```

Table A.17 shows additional square-table analyses of Table 8.6. First the data are entered as a 4×4 table, and the loglinear model fitted is quasi independence. The *qi* factor invokes

Table A.17 SAS code showing square-table analyses of Table 8.6.

```
-------------------------------------------------------------------------------
data square;
   input recycle  drive  qi  count @@;
datalines;
 1 1 1 12   1 2 5 43    1 3 5 163     1 4  5 233
 2 1 5  4   2 2 2 21    2 3 5  99     2 4  5 185
 3 1 5  4   3 2 5  8    3 3 3  77     3 4  5 230
 4 1 5  0   4 2 5  1    4 3 5  18     4 4  4 132
;
proc genmod; class drive recycle qi;
   model count = drive recycle qi / dist=poi link=log; * quasi indep;
data square2;
   input score below above @@; trials = below + above;
datalines;
 1 4 43   1 8 99    1 18 230    2 4 163    2 1 185    3 0 233
;
proc genmod data=square2;
   model above/trials = / dist=bin link=logit noint; * symmetry;
proc genmod data=square2;
   model above/trials = score / dist=bin link=logit noint; * quasi symmetry;
-------------------------------------------------------------------------------
```

the δ_i parameters. It takes a separate level for each cell on the main diagonal and a common value for all other cells. The bottom of Table A.17 fits logit models for the data entered in the form of pairs of cell counts (n_{ij}, n_{ji}). These six sets of binomial counts are labeled as *above* and *below* with reference to the main diagonal. The variable defined as *score* is the distance $(u_j - u_i) = j - i$. The first model is *symmetry* and the second is *ordinal quasi-symmetry*. Neither model contains an intercept (NOINT). The quasi-symmetry model can be fitted using the approach shown next for the equivalent Bradley–Terry model.

Table A.18 uses GENMOD for logistic fitting of the Bradley–Terry model to Table 8.8 by forming an artificial explanatory variable for each tennis player. For a given observation, the variable for player i is 1 if he wins, -1 if he loses, and 0 if he is not one of the players for that match. Each observation lists the number of wins for the player with the variate-level equal to 1 out of the number of matches (n) against the player with the variate-level equal to -1. The model has these artificial variates, one of which is redundant, as explanatory variables with no intercept term.

Table A.18 SAS code for fitting the Bradley–Terry model to tennis data in Table 8.8.

```
-------------------------------------------------------------------------------
data tennis;
input win n Djokovic Federer Murray Nadal Wawrinka;
datalines;
 9  15  1 -1  0  0  0
14  17  1  0 -1  0  0
...
 4   7  0  0  0  1 -1
;
proc genmod;
model win/n= Djokovic Federer Murray Nadal Wawrinka / dist=bin link=logit noint;
-------------------------------------------------------------------------------
```

Chapters 9–10: Marginal Models and Random Effects Models (GLMMs)

Table A.19 uses GENMOD for marginal modeling of Table 9.1 using GEE with an ungrouped data file. The REPEATED statement specifies the variable name (here, *case*) that identifies the subjects for each cluster. Possible working correlation structures are TYPE=EXCH for exchangeable, TYPE=AR for autoregressive, TYPE=INDEP for independence, and TYPE=UNSTR for unstructured. Output shows estimates and standard errors under the naive working correlation and incorporating the empirical correlation through the sandwich estimate. Alternatively, the working association structure in the binary case can use the log odds ratio (e.g., with LOGOR=EXCH for exchangeability). The TYPE3 option with the GEE approach provides score-type tests about effects.

Table A.19 SAS code for marginal and random effects modeling of opinions about legalized abortion data in Table 9.1.

```
-------------------------------------------------------------------------------
data abortion;
   input case  gender question  response ;  * response=1 is yes;
datalines;
       1     1        1         1
       1     1        2         1
       1     1        3         1

    1850     0        1         0
    1850     0        2         0
    1850     0        3         0
;
proc genmod descending;  class case question;
   model response = gender question / dist=bin link=logit type3;
   repeated subject=case / type=exch corrw;
proc glimmix method=quad(qpoints=1000) data=abortion; class question case;
   model response = question gender / dist=binomial link=logit solution;
   random intercept / subject=case;
-------------------------------------------------------------------------------
```

Table A.20 uses GENMOD to implement GEE with a cumulative logit model for Table 9.2.

Table A.20 SAS code for marginal and random effects modeling of insomnia data in Table 9.2.

```
-------------------------------------------------------------------------------
data sleep;
   input case treat occasion outcome;
datalines;
   1       1     0        1
   1       1     1        1
...
 239       0     0        4
 239       0     1        4
;
proc genmod data=sleep; class case;
 model outcome = treat occasion treat*occasion / dist=multinomial link=cumlogit;
    repeated subject=case / type=indep corrw;
proc glimmix  method=quad(qpoints=50) data=sleep; class case;
 model outcome=treat occasion treat*occasion / link=clogit dist=mult solution;
    random int / subject=case;
-------------------------------------------------------------------------------
```

PROC GLIMMIX and PROC NLMIXED can fit models that contain random effects, such as GLMMs. Table A.19 uses GLIMMIX to fit the GLMM to Table 9.1 on opinions about legalized abortion, setting the number of quadrature points (e.g., *qpoints* =). Table A.20 uses GLIMMIX for the insomnia clinical trial of Table 9.2. Table A.21 uses NLMIXED for Table 10.2 on basketball shooting.

Table A.21 SAS code for GLMM analyses of basketball data in Table 10.2.

```
-------------------------------------------------------------------------------
data basket;
input player $ y n;
datalines;
 Davis 32 39
....
 Gobert 11 14
;
proc nlmixed;
   eta = alpha + u;  p = exp(eta)/(1 + exp(eta));
   model y ~ binomial(n,p);
   random u ~ normal(0,sigma*sigma) subject=player;
   predict p out=new;
proc print data=new;
-------------------------------------------------------------------------------
```

Chapter 11: Non-Model-Based Classification and Clustering

PROC DISCRIM in SAS can perform discriminant analysis. Table A.22 does this for the horseshoe crab data (with y indicating whether a crab has at least one satellite). The statement *priors prop* sets the prior probabilities equal to the sample proportions, whereas *priors equal* sets them each to 0.50.

Table A.22 SAS code for discriminant analysis for horseshoe crab satellite data of Table 3.2.

```
-------------------------------------------------------------------------------
proc discrim data=crab crossvalidate;
   priors prop; class y;
   var width color;
-------------------------------------------------------------------------------
```

You can construct classification trees using PROC HPSPLIT. See

`https://support.sas.com/documentation/onlinedoc/stat/141/hpsplit.pdf`

PROC DISTANCE can assess similarities between pairs of variables, such as with simple matching. Then, PROC CLUSTER can perform a cluster analysis. See the SAS section at `www.stat.ufl.edu/~aa/cda/cda.html` for an example.

PROC GAM fits generalized additive models, as shown at the bottom of Table A.6. The lasso is available in PROC HPGENSELECT and can be implemented with PROC NLMIXED. See `support.sas.com/kb/60/240.html`.

A.3 `STATA` FOR CATEGORICAL DATA ANALYSIS

For `Stata`, basic support information is available at `www.stata.com/support`. Many Internet sites can help you learn how to use Stata, such as the many resources listed at

`www.stata.com/links/resources-for-learning-stata`

These tutorials and the discussion below show commands to use functions for various statistical analyses. Functions are case-sensitive. To get information about a function, use the `help` command, such as

```
help glm
```

for help about the `glm` function for fitting generalized linear models. Rather than entering commands in Stata, you can use the *Statistics* menu. Once you select a particular topic and subtopic, you get a dialog box in which you select the particular analyses you want.

Starting with version 15, Bayesian inference is available in Stata with the `bayes:` prefix before many functions, such as shown below for logistic regression. For details, see `www.stata.com/manuals/bayes.pdf`.

Chapters 1–2: Introduction and Contingency Tables

Using binomial counts, you can find a significance test (the score test) and various confidence intervals for a proportion. Here we illustrate with 9 successes in 10 trials and the score confidence interval (Wilson was the statistician who first proposed it):

```
cii proportions 10 9, wilson
prtesti 10 0.90 0.50 // provide n, sample prop., H0 prop.
```

Replace *wilson* by *wald* to obtain the Wald confidence interval, by *agresti* to obtain the Agresti–Coull interval, and by *jeffreys* to obtain the Jeffreys Bayesian posterior interval.

With the `tabulate` function (`tab` for short), you can construct contingency tables, find percentages in the conditional distributions (within-row relative frequencies), get expected frequencies for H_0: independence and the chi-squared statistic and its P-value, and conduct Fisher's exact test. For categorical variables x and y in an ungrouped data file, for instance:

```
tab x y, row expected chi2 exact
```

If you already have the cell counts, you can enter them by row. For Table 2.4 (in Section 2.4.4) on political party ID and gender, use

```
tabi 495 272 590 \ 330 265 498, row expected chi2 exact
```

To get standardized residuals, you currently must install a routine. Use the command

```
ssc install tab_chi
```

then followed (if you have the cell counts) by

```
tabchii 495 590 272 \ 330 498 265, adjust
```

to get the standardized (adjusted) residuals.

Stata can find the Wald confidence interval for an odds ratio (originally derived by Woolf) with a command (such as for Table 2.7)

```
csi 3 1 1 3 , or woolf
```

Chapters 3–5: Generalized Linear Models and Logistic Regression

Stata can fit generalized linear models with the `glm` function. For instance, you can fit a logistic regression model by treating it as a GLM for a binomial distribution with logit link, such as by

```
glm y x, family(binomial) link(logit)
```

for a response y and explanatory variable x in an ungrouped data file. To fit the linear probability model, substitute *identity* for the link. For probit models, substitute *probit* as the link.

Stata can also fit logistic regression models with the `logit` function, for which the standard output is the model parameter estimates, or the `logistic` function, for which the standard output is the odds ratios obtained by exponentiating the estimates. For example, for a binary response variable with three explanatory variables in an ungrouped data file:

```
logit y x1 x2 x3
```

Adding the *or* option to this command requests the odds ratio form of estimate.

If the data are counts in a contingency table, and each row of the data file has a value for each explanatory variable, the 0 or the 1 value for y, and a variable (say, called *count*) containing the cell counts, you can enter:

```
logit y x1 x2 x3 [fweight = count]
```

Here, *fweight* = *count* indicates that the data file has data grouped according to the variable called *count*.

To perform a likelihood-ratio test about an individual explanatory variable, store the results for the full model, fit the simpler model without that variable, and then request the likelihood-ratio test comparing the models. For example, to test the effect of *x1* in a model that also has predictor *x2*,

```
------------------------------------------------------------------------------
logit y x1 x2 [fweight = count]
estimates store full
logit y x2 [fweight = count]
lrtest full
------------------------------------------------------------------------------
```

The profile likelihood confidence interval is available with the *pllf* function. A command such as

```
test x1
```

conducts the Wald test about an explanatory variable.

For Bayesian fitting, starting with version 15 use the `Bayes` prefix, such as

```
bayes, normalprior(10): logit y x1 x2 x3
```

to use normal priors with standard deviation 10 (the default is 100). This also works with the `glm` function and its links, such as

```
bayes: glm y x1 x2 x3, family(binomial) link(probit)
```

For conditional logistic regression, use a command such as

```
clogit y x1, group(x2)
```

to condition on $x2$ in analyzing the effect of $x1$.

Chapters 6–7: Multicategory Logit Models and Loglinear Models

Stata fits the baseline-category logit model with the `mlogit` (multinomial logit) function, such as

```
mlogit y x1 x2, base(3)
```

where *base(3)* indicates the baseline category for the logits. If the data file contains grouped data (i.e., cell counts in the response categories), such as columns labeled *x1*, *x2*, *y* (giving the response category), and *count*, fit the model with the command

```
mlogit y x1 x2  [fweight = count], base(3)
```

Stata fits the cumulative logit model with a proportional odds structure using the `ologit` (ordinal logit) function, such as

```
ologit y x1 x2
```

with the parameterization (6.6) based on the latent variable model. If the data file contains grouped data, such as columns labeled *x* (say, a 1/0 indicator), *y* (giving the response category), and *count*, fit the model with the command

```
ologit y x [fweight = count]
```

For Bayesian fitting of multinomial models, starting with version 15 use the `Bayes` prefix, such as

```
bayes, normalprior(10): ologit y x1 x2
```

for cumulative logit models. Replace `ologit` by `mlogit` for baseline-category logit models.

Stata can fit loglinear models by regarding them as GLMs with a response count having a Poisson distribution, using the log link. For example, for a three-way contingency table constructed from a data file with three columns of levels for the variables (say, x, y, and z) and a column of cell counts (labelled *count*), fit the homogeneous association model by

```
glm count i.x i.y i.z i.x#i.y i.x#i.z i.y#i.z, family(poisson) link(log)
```

which creates indicator variables for the classification variables.

Chapters 8–11: Correlated Observations, Advanced Methods

When $y1$ and $y2$ in a data file are binary, you can get McNemar's test using

```
mcc y1 y2
```

Using the summary counts in the contingency table that cross-classifies $y1$ and $y2$, you can get McNemar's test (e.g., for the example in Section 8.1.1) using

```
mcci 227 132 107 678
```

You can obtain GEE fitting in Stata with the `xtgee` function. For example, using the `Abortion` data at the text website:

```
xtgee response gender i.situation, family(binomial) link(logit) corr(exchangeable)
```

For random effects models, the `GLLAMM` module is very powerful. You can also fit logistic mixed effects models (GLMMs), including multilevel models, with the `melogit` function, as described at `www.stata.com/manuals/me.pdf`; for example,

```
melogit response gender i.situation || person: intpoints(100)
```

which uses *intpoints* to specify the number of quadrature points.

For linear discriminant analysis, use the `discrim` function. For the horseshoe crab data analyzed in Section 11.1.2,

```
discrim lda width color, group(y) priors(proportional) lootable
```

where *lootable* requests the classification table after leave-one-out cross-validation. The *priors*(*proportional*) option uses the sample proportions in the two categories for the prior probabilities for prediction. If you omit this, the prior probabilities are both 0.50.

Cluster analysis is available with the `cluster` function, such as

```
cluster averagelinkage e1 e2 e3 e4 e5 e6 e7 e8 e9 e10, measure(matching)
cluster dendrogram
```

with simple similarity matching for the `Elections` and `Elections2` data files at the text website.

A.4 SPSS FOR CATEGORICAL DATA ANALYSIS

SPSS has a windows-with-menus structure that makes requesting statistical procedures simple. In the *Variable View* of the *Data Editor* window, SPSS should identify quantitative variables as NUMERIC and categorical variables (with labels for the categories) as STRING. You can redefine names and characteristics for each variable. In the Measure column, make sure SPSS has not inappropriately labeled a variable as NOMINAL that should be SCALE (interval) or ORDINAL.

When you select a statistical procedure from the ANALYZE menu on the *Data Editor*, a *dialog box* opens that shows you the source variables in your data set. When you click on the procedure you want, results show in the output window. For many procedures, you can click on *Options* and an additional *subdialog box* will open that displays extra options.

Chapters 1–2: Introduction and Contingency Tables

On the ANALYZE menu, the DESCRIPTIVE STATISTICS option has a suboption called CROSSTABS, which provides several methods for contingency tables. After identifying the row and column variables in CROSSTABS, clicking on STATISTICS provides a wide variety of options, including the Pearson and likelihood-ratio chi-squared tests. You can request Fisher's exact test by clicking on EXACT in the CROSSTABS dialog box and selecting the exact test.

In CROSSTABS, clicking on CELLS provides options for displaying observed and expected frequencies, as well as the standardized residuals, labeled as *Adjusted standardized*. In CROSSTABS, the options in STATISTICS include measures of association. One option, labeled *Risk*, provides as output for 2×2 tables the odds ratio and its confidence interval.

Suppose you enter the data as cell counts for the various combinations of the two variables, rather than as ungrouped responses on the variables for individual subjects; for instance, perhaps you call COUNT the variable that contains these counts. Then, select the WEIGHT CASES option on the DATA menu in the *Data Editor* window and instruct SPSS to weight cases by COUNT.

Chapters 3–5: Generalized Linear Models and Logistic Regression

You can fit GLMs with the GENERALIZED LINEAR MODELS option and suboption. For example, to fit logistic regression models, you pick the Binary response and Binary logistic model. With the *Response* tab you can also enter the data as the number of successes out of a certain number of trials, for grouped data. For example, suppose in one column you have the number of successes at each particular setting of predictors and in a separate column you have the sample size that the number of successes is based on. Then, you identify the dependent variable as the variable listing the number of successes, you click the box "Variable represents binary response or number of events," and then "Number of events occurring in a set of trials," entering the variable listing the sample sizes as the "Trials variable."

To fit logistic regression models, you can also select the REGRESSION option and the BINARY LOGISTIC suboption. In the LOGISTIC REGRESSION dialog box, identify the binary response (dependent) variable and the explanatory predictors (covariates). Identify the explanatory variables that are categorical and for which you want indicator variables by clicking on Categorical and declaring such a covariate to be a Categorical Covariate in the LOGISTIC REGRESSION: DEFINE CATEGORICAL VARIABLES dialog box. Highlight the categorical covariate and under Change Contrast you will see several options for setting up indicator variables. The *Simple* contrast constructs them with the final category as the baseline.

In the LOGISTIC REGRESSION dialog box, click on *Method* for stepwise model selection procedures, such as backward and forward selection. Click on *Save* to save predicted probabilities, measures of influence such as standardized residuals. Click on *Options* to

open a dialog box that contains an option to construct confidence intervals for exponentiated parameters.

Chapters 6–7: Multicategory Logit Models and Loglinear Models

Select the MULTINOMIAL LOGISTIC suboption for a baseline-category logit model. Click on *Statistics* and check Likelihood-ratio tests under Parameters to obtain likelihood-ratio tests for effects of the predictors. Choose the REGRESSION option and then the ORDINAL suboption for a cumulative logit model, which is fitted with the latent variable parameterization. This model is also available under the GENERALIZED LINEAR MODELS option.

For loglinear models, use the LOGLINEAR option with GENERAL suboption. (You can also select a Poisson loglinear model with the GENERALIZED LINEAR MODELS option.) You enter the factors for the model. The default is the saturated model, so click on *Model* and select a *Custom* model. Enter the factors as terms in a customized (unsaturated) model and then select additional interaction effects. Click on *Options* to show options for displaying observed and expected frequencies and adjusted residuals. When the data file contains the data as cell counts for the various combinations of factors rather than as ungrouped responses for individual subjects, weight each cell by the cell count using the WEIGHT CASES option in the DATA menu.

Chapters 8–11: Correlated Observations, Advanced Methods

McNemar's test is available with the NONPARAMETRIC TESTS option. For details, see

`www.spss-tutorials.com/spss-mcnemar-test`

You can also fit square-table loglinear models. For quasi symmetry, see

`www.ibm.com/support/knowledgecenter/SSLVMB_23.0.0/spss/tutorials/genlog_debate_howto_02.html`

and for quasi-independence, see

`www.ibm.com/support/knowledgecenter/en/SSLVMB_23.0.0/spss/tutorials/genlog_debate_howto_04.html`

For the kappa measure of agreement, specify the DESCRIPTIVE STATISTICS option and the CROSSTABS suboption. Specify the raters as the row and column variables and click the STATISTICS menu and check the KAPPA box. You can fit the Bradley–Terry model as a quasi-symmetry model.

For GEE methods, use the GENERALIZED LINEAR MODELS with GENERALIZED ESTIMATING EQUATIONS suboption. You specify the distribution, link function, response variable, factors and covariates, and specify the model effects and the subject variable on which repeated observations occur.

For linear discriminant analysis, specify the CLASSIFY option and DISCRIMINANT suboption, and select the group outcome and the explanatory variables. For cluster analysis, specify the HIERARCHICAL CLUSTER suboption. The DECISION TREES module can construct classification trees.

BRIEF SOLUTIONS TO ODD-NUMBERED EXERCISES

CHAPTER 1

1.1 Response variables: (a) Attitude toward gun control, (b) Heart disease, (c) Vote for President.

1.3 a. Binomial, $n = 100$, $\pi = 0.25$. b. $\mu = n\pi = 25$ and $\sigma = \sqrt{n\pi(1-\pi)} = 4.33$. 50 correct responses is surprising, because 50 is $z = (50 - 25)/4.33 = 5.8$ standard deviations above the mean.

1.5 Uniform prior has $\alpha = \beta = 1$, Bayes estimator $= (y + 1)/(n + 2) = 0.25$.

1.7 a. $(5/6)^6$. b. Obtain $Y = y$ when you have $y - 1$ successes and then a failure.

1.9 a. Let $\pi =$ population proportion obtaining greater relief with new analgesic. For $H_0 : \pi = 0.50$, $z = 2.00$, P-value $= 0.046$.

b. Wald CI is (0.504, 0.696), score CI is (0.502, 0.691).

1.11 Wald CI is $0.86 \pm 1.96(0.0102)$, or (0.84, 0.88).

1.13 a. $(1 - \pi_0)^{25}$ is binomial probability of $y = 0$ in $n = 25$ trials. The maximum of $(1 - \pi)^{25}$ occurs at $\pi = 0.0$.

b. $2\log(\ell_1/\ell_0) = 2\log[1.0/(0.50)^{25}] = 34.7$, P-value < 0.0001.

c. $2\log(\ell_1/\ell_0) = -2\log[1.0/(0.926)^{25}] = 3.84$. With $df = 1$, chi-squared P-value $= 0.05$.

1.15 a. $\sigma(\hat{\pi})$ equals binomial standard deviation $\sqrt{n\pi(1-\pi)}$ divided by n.

b. $\sigma(\hat{\pi})$ takes maximum at $\pi = 0.50$ and minimum at $\pi = 0$ and 1.

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

1.17 Posterior mean is $(y + 0.5)/(n + 1) = 0.5/26 = 0.0192$, posterior interval is found in R using

```
> qbeta(0.025, 0.5, 25.5); qbeta(0.975, 0.5, 25.5)
```

Posterior $P(\pi < 0.50)$ is found using `pbeta(0.50, 0.5, 25.5) = 1`.

1.19 Score: With repeated random samples of size 25, in the long run 95% of such CIs would contain an actual value of π. Bayes: After seeing the data, we conclude that the probability is 0.95 that π falls between 0.00002 and 0.0947.

CHAPTER 2

2.1 a. $P(- \mid C) = 1/4$, $P(+ \mid \bar{C}) = 1/10$.
b. Sensitivity $= P(+ \mid C) = 1 - P(- \mid C) = 3/4$; specificity $= P(- \mid \bar{C}) = 1 - P(+ \mid \bar{C}) = 9/10$.
c. $P(C, +) = P(+ \mid C)P(C) = (3/4)(0.04) = 0.03$,
$P(C, -) = P(- \mid C)P(C) = 0.01, P(\bar{C}, +) = 0.096, P(\bar{C}, -) = 0.864$.
d. $P(+) = 0.126, P(-) = 0.874$; $P(C \mid +) = 0.03/0.126 = 0.238$.

2.3 a. 0.000061. b. 62.4/1.3 = 48. Relative risk.

2.5 a. Relative risk. b. (i) $\pi_1 = 0.55\pi_2$, so $\pi_1/\pi_2 = 0.55$. (ii) 1/0.55 = 1.82.

2.7 a. Quoted interpretation is that of relative risk.
b. Proportion = 0.744 for females, 0.203 for males. $RR = 0.744/0.203 = 3.7$.

2.9 Difference of proportions: Lung cancer, 0.00130; Heart disease, 0.00256. Cigarette smoking seems more highly associated with heart disease.
Odds ratio: Lung cancer, 14.02; Heart disease, 1.62; e.g., the odds of dying from lung cancer are estimated to be 14.02 times higher for smokers than for nonsmokers. The difference of proportions describes excess deaths due to smoking. That is, if $N =$ no. of smokers in population, we predict there would be $0.00130N$ fewer deaths per year from lung cancer if they had never smoked, and $0.00256N$ fewer deaths per year from heart disease. Thus elimination of cigarette smoking would have the biggest impact on deaths due to heart disease.

2.11 Estimated odds ratio = 219.86; Wald CI is (140.9, 343.1).

2.13 $X^2 = 35.7$, $df = 1$, P-value < 0.0001. Belief in an afterlife is associated with gender.

2.15 a. Responses = cancer (yes, no) and colorectal cancer (yes, no); explanatory = regular aspirin use (yes, no). For those taking aspirin regularly, we are 95% confident that the population proportion who got colorectal cancer was between 0.75 and 0.88 times the population proportion of colorectal cancer occurrences for those who did not take aspirin regularly.
b. (i) Significant, because CI does not contain 1.0; (ii) weak, because plausible values are very close to 1.0.

2.17 a. $G^2 = 213.9$, $X^2 = 184.3$, $df = 2$ ($P < 0.0001$).

b. Standardized residuals of -11.97 for white Democrats and -13.00 for black Republicans show extremely strong evidence of fewer people in these cells than if party ID were independent of race. Standardized residuals of 11.97 for black Democrats and 13.00 for white Republicans show extremely strong evidence of more people in these cells than expected.

c. $G^2 = 213.6$ for comparing races on (Democrat, Republican) choice and $G^2 = 0.3$ for comparing races on (Democrat + Republican, Independent) choice.

2.19 Extremely strong evidence of tendency for those with less than high school education to be fundamentalist and those with bachelor degree or higher to be liberal in religious beliefs. Could alternatively use test based on correlation, since variables are ordinal.

2.21 a. $X^2 = 11.5, df = 6, P = 0.24$; nominal test with ordinal data. Residuals suggest tendency for job satisfaction to be higher when income is higher.

b. Ordinal test gives $M^2 = 7.04$, $df = 1$, $P = 0.008$, and much stronger evidence of an association, using ordinality to detect a trend.

2.23 a. The P-values are 0.38 and 0.64. For either alternative, it is plausible that control of cancer is independent of treatment used.

b. Two-sided mid P-value $= 0.486$, CI is (0.275, 19.155). The sum of the two one-sided mid P-values is 1.0 and the null expected value is 0.50. Inference is less conservative using the mid P-value.

2.25 a. Sample conditional odds ratios are 0.680 for white victims and 0.00 for black victims.

b. 1.173. Yes, Simpson's paradox occurs, because marginal association is in a different direction than partial associations; reason for switch in association is the same as in the text example.

2.27 0.18 for males and 0.32 for females.

2.29 a. beta(1230.5, 357.5) and beta(859.5, 413.5) have posterior means 0.775 and 0.675.

b. Posterior intervals (0.067, 0.133) for difference of proportions and (1.40, 1.96) for odds ratio; posterior probability 1.0.

CHAPTER 3

3.1 The link function determines the function of the mean that is predicted by the linear predictor in a GLM. The identity link is not often used for binomial probabilities, because probabilities must fall between 0 and 1, whereas straight lines provide predictions that can be any real number.

3.3 a. $\hat{P}(Y = 1) = 0.00255 + 0.00109(\text{alcohol})$.

b. Estimated probability of malformation increases from 0.00255 at $x = 0$ to 0.01018 at $x = 7$. Relative risk $= 0.01018/0.00255 = 4.0$.

c. $\hat{P}(Y = 1) = 0.00260 + 0.00050(\text{alcohol})$. The estimated probabilities are 0.0026 and 0.0046, with relative risk 1.8.

d. Linear probability model has $\hat{P}(Y = 1) = 0.00263 + 0.00067$(alcohol). The estimated probabilities are 0.0026 and 0.0073, with relative risk 2.8.

3.5 a. With least squares, $\hat{P}(Y = 1) = -0.145 + 0.323$(weight); at weight = 5.2, $\hat{P}(Y = 1) = 1.53$, much higher than the upper bound of 1.0 for a probability.

b. logit$[\hat{P}(Y = 1)] = -3.695 + 1.815$(weight); at 5.2 kg, predicted logit = 5.74 and $\log$(0.9968/0.0032) = 5.74.

3.7 a. The fitted linear predictor for the logistic regression model is (i) $-3.777 + 0.327x$, (ii) $-3.777 + 0.655x$, (iii) $-4.432 + 0.655x$. Slope depends on distance between scores; doubling distance halves the slope estimate. Fitted values are identical for any linear transformation, (0.022, 0.042, 0.078, 0.140).

b. The ML logistic model fit has fitted values (0.023, 0.038, 0.102, 0.128), compared to (0.021, 0.044, 0.093, 0.132) with the scores (0, 2, 4, 5). The substantive results are the same.

3.9 a. logit$[\hat{P}(Y = 1)] = -3.518 + 0.1054x$. Since $\hat{\beta} = 0.1054 > 0$, the estimated probability of possessing a travel credit card increases as annual income increases.

b. Substituting $x = 33.4$ gives an estimated logit of 0 and thus an estimated probability of 0.50.

3.11 $\log(\hat{\mu}) = 1.609 + 0.588x$. $\exp(\hat{\beta}) = \hat{\mu}_B/\hat{\mu}_A = 1.80$.

3.13 a. $\log(\hat{\mu}) = -0.428 + 0.589$(weight); 2.74.

b. $0.589 \pm 1.96(0.065) =$ (0.462, 0.717); CI for multiplicative effect on mean is (1.59, 2.05).

c. Wald $z^2 = (0.589/0.065)^2 = 82.2$, LR statistic = 71.9, $df = 1$; strong evidence of positive effect.

3.15 a. T, b. F.

CHAPTER 4

4.1 a. $\hat{\pi} = 0.50$ at $-\hat{\alpha}/\hat{\beta} = 3.7771/0.1449 = 26$.

b. $e^{\hat{\beta}} = e^{0.1449} = 1.16$.

c. The lower quartile and upper quartile for LI are 14 and 28; $\hat{\pi}$ increases by 0.42, from 0.15 to 0.57, between those values.

d. At LI = 8, $\hat{\pi} = 0.068$, rate of change = 0.1449(0.068)(0.932) = 0.009.

e. Average marginal effect = 0.0226.

4.3

LI	Cases	Remissions
8	2	0
10	2	0
12	3	0
...		
38	3	2

ML estimates are the same, but deviance is not, because the number of parameters in the saturated model is 27 for ungrouped data but 14 for grouped data.

4.5 a. $\text{logit}(\hat{\pi}) = 15.043 - 0.232x$.

b. At temperature = 31, $\hat{\pi} = 0.9996$.

c. $\hat{\pi} = 0.50$ at $x = 64.8$ and $\hat{\pi} > 0.50$ at $x < 64.8$. At $x = 64.8$, $\hat{\pi}$ decreases at rate 0.058.

d. Estimated odds of thermal distress multiply by $\exp(-0.232) = 0.79$ for each 1-degree increase in temperature.

e. Wald statistic $z^2 = 4.6$ ($P = 0.03$) and LR statistic = 7.95 ($df = 1$, $P = 0.005$).

4.7 a. $\text{logit}(\hat{\pi}) = -0.573 + 0.0043(\text{age})$. LR statistic = 0.55, Wald statistic = 0.54, $df = 1$; no evidence of age effect.

b. Age values are more disperse when kyphosis absent.

c. $\text{logit}(\hat{\pi}) = -3.035 + 0.0558(\text{age}) - 0.0003(\text{age})^2$. LR statistic for $(\text{age})^2$ term equals 6.3 ($df = 1$), showing strong evidence of effect.

4.9 a. $\text{logit}(\hat{\pi}) = -0.76 + 1.86c_1 + 1.74c_2 + 1.13c_3$. The estimated odds a medium-light crab has a satellite are $e^{1.86} = 6.4$ times estimated odds a dark crab has a satellite.

b. LR statistic = 13.7, $df = 3$, P-value = 0.003.

c. For color scores 1,2,3,4, $\text{logit}(\hat{\pi}) = 2.36 - 0.71c$, the LR statistic = 12.5, $df = 1$, P-value = 0.0004.

d. Power advantage of focusing test on $df = 1$, but may not be linear trend for color effect.

e. 0.954, -0.412; a standard deviation increase in weight has more than double the effect of a standard deviation increase in color, adjusting for the other variable.

4.11 a. Predicted logit $-3.596 - 0.868d + 2.404v$, with $d = v = 1$ for white race. For example, for a given defendant's race, the odds of the death penalty when the victims were white are $e^{2.4044} = 11.1$ times the odds when the victims were black. Most likely are black defendants with white victims.

b. For a given defendant's race, the odds of the death penalty when the victims were white are estimated to be between 3.7 and 41.2 times the odds when the victims were black. Wald statistic = 5.6, LR statistic = 5.0, each with $df = 1$. P-value = 0.025 for LR statistic.

4.13 The death penalty was most likely for black defendants with white victims. The estimated odds of a black defendant receiving the death penalty were $e^{1.1886} = 3.28$ times the odds for a white defendant, controlling for the victim's race. The effect of the victim's race was highly statistically significant.

4.15 a. $\text{logit}(\hat{\pi}) = \hat{\alpha} + 1.40x_1 + 0.32x_2 + 1.76x_3 + 1.17x_4$, where $x_1 = 1$ for educated and 0 for non-educated, $x_2 = 1$ for males and 0 for females, $x_3 = 1$ for high SES and 0 for low SES, and $x_4 =$ lifetime no. of partners.

b. CI corresponds to one for log odds ratio of (0.207, 2.556); 1.38 is midpoint of CI, suggesting it may be estimated log odds ratio, in which case $\exp(1.38) = 3.98 =$ estimated odds ratio.

4.17 a. $r = 1$: $\text{logit}(\hat{\pi}) = -6.7 + 0.1a + 1.4s$. $r = 0$: $\text{logit}(\hat{\pi}) = -7.0 + 0.1a + 1.2s$. Conditional odds ratio $= \exp(1.4) = 4.1$ for blacks and $\exp(1.2) = 3.3$ for whites. Coefficient of cross-product term, 0.22, is difference between log odds ratios 1.4 and 1.2.

b. The coefficient of s of 1.2 is log odds ratio between y and s when $r = 0$ (whites), in which case interaction does not enter the equation. P-value of $P < 0.01$ for smoking represents result of test that log odds ratio between y and s for whites $= 0$.

4.19 a. Derive the four equations from the overall equation $\text{logit}(\hat{\pi}) = -5.854 + 4.101c_1 - 4.186c_2 - 15.66c_3 + 0.200x - 0.094(c_1 \times x) + 0.218(c_2 \times x) + 0.658(c_3 \times x)$.

b. LR statistic $= 4.4$ ($df = 3$), $P = 0.22$. Correlation $= 0.452$ for simpler model and 0.472 allowing interaction. The simpler model fits nearly as well and is adequate according to the test.

4.21 We predict $y = 1$ whenever $\hat{\pi} > 0.50$. Of the cases with $y = 1$, they are classified as correct or incorrect according to whether $\hat{\pi} > 0.50$ or not. Of the cases with $y = 0$, they are classified as correct or incorrect according to whether $\hat{\pi} < 0.50$ or not. Sensitivity $= 96/111 = 0.865$, specificity $= 31/62 = 0.50$.

CHAPTER 5

5.1 a. LR statistic $= 32.9$ ($df = 2$), $P < 0.0001$.

b. LR statistics are 1.56 for weight and 2.85 for width, with P-values of 0.21 and 0.09. The predictors are highly correlated (correlation $= 0.887$), so multicollinearity occurs.

5.5 a. Deviance $= 11.1$, $df = 11$, so no evidence of lack of fit. Model is adequate. Take out JP term, as it is the least significant.

b. The model with only the four main effect terms, which has the smallest AIC.

5.7 a. Deviance $= 1.38$, $df = 1$, has P-value $= 0.24$. The model fits adequately.

b. The fitted values are close for the two race groups that used AZT and also close for the two race groups that did not use AZT, which reflects the lack of significance for race in the model.

c. Standardized residuals have approximate standard normal distributions (when model holds) and appropriately reflect residual df values (i.e., only one is nonredundant).

5.9 a. $\text{logit}(\pi) = \alpha + \beta_2 d_2 + \cdots + \beta_6 d_6$, where $d_i = 1$ for department i and $d_i = 0$ otherwise. The model fits poorly, with residual deviance $= 21.7$ with $df = 6$. The standardized residuals for the number of females who were admitted are (4.15, 0.50, -0.87, 0.55, -1.00, 0.62). The only lack of fit is in Department 1, where more females were admitted than expected. For males, the standardized residual $= -4.15$, so fewer males were admitted than expected if the model lacking gender effect truly holds.

b. Males apply in relatively greater numbers to departments that have relatively higher proportions of acceptances.

5.11 More complex models containing additional explanatory variables or interaction terms do not provide an improved fit. A residual analysis is not informative because the data are ungrouped rather than a nonsparse contingency table.

5.13 Any model with *affirm* as an explanatory variable, because of those who did not support affirmative action, 0 were vegetarians.

5.15 a. Some centers have successes but some do not have any.
b. Depends on software used, but the actual ML estimates are both $-\infty$.
c. $\hat{\beta} = 1.55$ ($SE = 0.70$).

5.17 The linear probability model, fitted with least squares, estimates that adjusting for duration, the estimated probability of a sore throat is 0.32 lower for the tracheal tube than the laryngeal mask. The probit model estimates that the latent distribution on the sore throat response is shifted 0.93 standard deviations lower for the tracheal tube than the laryngeal mask.

5.19 a. $z_{\alpha/2} = 1.645$, $z_\beta = 0.842$, and $n_1 = n_2 = 229$.

5.21 a. F, b. T

CHAPTER 6

6.1 a. $\log(\hat{\pi}_R/\hat{\pi}_D) = -2.3 + 0.5x$. Estimated odds of preferring Republicans over Democrats increase by 65% for every \$10,000 increase.
b. $\hat{\pi}_R > \hat{\pi}_D$ when annual income >\$46,000.
c. $\hat{\pi}_I = 1/[1 + \exp(3.3 - 0.2x) + \exp(1 + 0.3x)]$.

6.3 Here are effects for prediction equations, when the length indicator is 1 for the small length and 0 otherwise, with SE values in parentheses.

Logit	Inter.	Length $\leq$2.3	Hancock	Oklawaha	Trafford
$\log(\pi_I/\pi_F)$	-1.55	1.46(0.40)	-1.66(0.61)	0.94(0.47)	1.12(0.49)
$\log(\pi_R/\pi_F)$	-3.31	-0.35(0.58)	1.24(1.19)	2.46(1.12)	2.94(1.12)
$\log(\pi_B/\pi_F)$	-2.09	-0.63(0.64)	0.70(0.78)	-0.65(1.20)	1.09(0.84)
$\log(\pi_O/\pi_F)$	-1.90	0.33(0.45)	0.83(0.56)	0.01(0.78)	1.52(0.62)

Since the effect estimate $1.46 > 0$, invertebrates are relatively more likely than fish for the small length of alligators.

6.5 a. Job satisfaction tends to increase at higher x_1 and lower x_2 and x_3.
b. $x_1 = 4$ and $x_2 = x_3 = 1$.

6.7 a. This model takes advantage of response being ordinal and is more parsimonious.
b. Deviance = 3.25, $df = 3$, P-value = 0.36, so the model fits adequately. It has two cumulative probabilities to model, and hence two intercept parameters. The proportional odds structure has the same predictor effect for each cumulative probability, so only one effect is reported for income.
c. The likelihood-ratio statistic equals $4.13 - 3.25 = 0.89$ with $df = 1$, and a P-value of 0.35. It is plausible that income has no effect on marital happiness. Since $\hat{\beta} = -0.112 < 0$, the estimated odds of being at low end of scale (less happy) decrease as income increases.

6.9 a. $c = 5$; (i) religion = Other, (ii) religion = Protestant.

b. $e^{-1.27-(-1.22)} = 0.95$; i.e., the estimated odds that a Protestant falls in relatively more liberal categories (rather than more conservative categories) is 0.95 times the estimated odds for a Catholic.

6.11 a. For high SES, $\hat{P}(Y = 1)$ changes from 0.696 to 0.115 as life events increases from its minimum to maximum values; for low SES, it changes from 0.430 to 0.041.

b. $R^2 = 0.224$, $R = 0.473$.

c. The life events and SES effects change only to 0.302 and -1.263. If the model holds for an underlying logistic latent variable, the model holds with the same effect value for every way of defining the outcome categories.

d.

```
------------------------------------------------------------------------
               Mean      SD  Median     2.5%  97.5%  MC error/SD
logOR(life)   0.249  0.0997   0.247   0.0571  0.448      0.00628
logOR(ses)   -0.771  0.5120  -0.768  -1.7800  0.228      0.00165
------------------------------------------------------------------------
```

The posterior interval for the cumulative log odds ratio for the effect of SES is then $(-1.780, 0.228)$ and the posterior $P(\beta_2 > 0) = 0.065$. With prior beliefs that effects are not large, posterior results exhibit considerable shrinkage toward the no effect value of 0.

6.13 For $\hat{\beta}_1 = 3.634$ for the latent variable model parameterization, $\hat{P}(Y_2^* > Y_1^*) = \exp(\hat{\beta}_1/\sqrt{2})/[1 + \exp(\hat{\beta}_1/\sqrt{2})] = 0.93$. For either gender, a strong Republican is very likely to be more conservative than a strong Democrat.

6.15 a. Using a cumulative logit model with indicator variables for race (1 = black, 0 = white) and for gender (1 = female, 0 = male), we get race effect -0.266 ($SE = 0.200$) and gender effect 0.379 ($SE = 0.141$). The model fits adequately (e.g., deviance = 1.14, $df = 4$).

b. Using the indicator variables from (a), the estimated effects for the log odds of being in the lower of two adjacent belief categories are -0.163 for race ($SE = 0.117$) and 0.219 for gender ($SE = 0.083$).

6.17 The sequential logits model the probability of a dead fetus, using $\log[\pi_1/(\pi_2 + \pi_3)]$, and the conditional probability of a malformed fetus, given that the fetus was live, using $\log(\pi_2/\pi_3)$. The estimated effect of concentration level is 0.0064 ($SE = 0.0004$) for the first logit and 0.0174 ($SE = 0.0012$) for the second logit. In each case, the less desirable outcome is more likely as the concentration level increases. The simpler model with proportional odds structure fits more poorly (increase in deviance = 117.64, $df = 1$).

6.19 Cumulative logit model with main effects of gender, location, and seat-belt has estimates 0.545, -0.773, and 0.824. For example, for those wearing a seat-belt, the estimated odds that the response is below any particular level of injury and are $e^{0.824} = 2.3$ times the estimated odds for those not wearing seat-belts.

CHAPTER 7

7.1 a. It is plausible that belief in afterlife is independent of gender.

b. $\hat{\lambda}_1^Y = 0.0$, $\hat{\lambda}_2^Y = 1.416$. Given gender, estimated odds of believing in afterlife equal $e^{1.416} = 4.1$ (i.e, 4.1 people believe in afterlife for every 1 who does not).

c. The estimated odds that a female believes in the afterlife are $e^{0.1368} = 1.15$ times the estimated odds for a male.

d. Output for the independence model is:

```
---------------------------------------------------------------------
                     Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)           4.05242     0.07309    55.44    <2e-16
factor(race)other    -1.05209     0.11075    -9.50    <2e-16
factor(race)white     1.64927     0.06152    26.81    <2e-16
factor(postlife)yes   1.49846     0.05697    26.30    <2e-16
---
Residual deviance:    0.35649  on 2  degrees of freedom
---------------------------------------------------------------------
```

$\hat{\lambda}_1^Y - \hat{\lambda}_2^Y = 1.498$; assuming the independence model, odds of belief in afterlife are $e^{1.498} = 4.5$ for each race.

7.3 a. Let $S =$ safety equipment, $E =$ whether ejected, $I =$ injury. Then, deviance for the homogeneous association model (SE, SI, EI) is 2.85, $df = 1$. Any simpler model has deviance > 1000, so it seems there is an association for each pair of variables and that association can be regarded as the same at each level of the third variable. The estimated conditional odds ratios are 0.091 for S and E (i.e., wearers of seat-belts are much less likely to be ejected), 5.57 for S and I, and 0.061 for E and I.

7.5 a. Injury has estimated conditional odds ratios 0.58 with gender, 2.13 with location, and 0.44 with seat-belt use. The odds of no injury for females are estimated to be 0.58 times the odds of no injury for males (controlling for L and S); that is, females are more likely to be injured. Similarly, the odds of no injury for urban location are estimated to be 2.13 times the odds for rural location, so injury is more likely at a rural location, and the odds of no injury for no seat-belt use are estimated to be 0.44 times the odds for seat-belt use, so injury is more likely for no seat-belt use, other things being fixed. Since there is no interaction for this model, overall the most likely case for injury is therefore females not wearing seat-belts in rural locations.

b. You could model the choice about whether to wear a seat-belt, using various predictors; G and L are possible predictors for S. Then, conditional on this choice, you could model the injury outcome, using G, L, and S as predictors. S is a response variable in the first analysis and an explanatory variable in the second analysis.

7.7 With two-factor terms, deviance $= 31.7$, $df = 48$. With three-factor terms, deviance $= 8.5$, $df = 16$; the change is $31.7 - 8.5 = 23.1$, $df = 32$, not a significant improvement. The estimated conditional odds ratios are 8.5 for E and H, 2.4

for C and L, 6.5 for H and L, 0.8 for C and H, 0.9 for E and L, 3.3 for C and E. The simpler model deleting the CH and EL terms, which show weak associations, fits well (deviance $= 39.4$, $df = 56$).

7.9 A logistic model is more appropriate when one of the variables is a response and the others are explanatory. A loglinear model may be more appropriate when at least two of the variables are response variables; one can use the model to describe the association between the responses and the effects of the explanatory variables on the responses.

7.11 b. (ii) For multiway collapsibility conditions, identify set $B = \{E, L\}$ and sets A and C each to be one of the other variables.

7.13 c. From the log of formula (7.3), the log odds ratio is proportional to the product of the distance between scores. A column distance of 2 gives a product that is double the value for a column distance of 1. One odds ratio equals the other one squared.

7.15 a. Deviance is 29.07. The deviance decrease of 14.86 on df $= 12$ is not a significantly better fit.

b. Deviance is 45.80. The deviance increase of 1.88 on df $= 2$ is not significant, suggesting that histology is not needed in the model.

7.17 a. No. There is clear evidence of substantial overdispersion, because the sample variances are so much larger than the means and the population mean and variance are equal for the Poisson.

b. Because the Poisson model does not take into account the overdispersion, hence giving an unrealistically small SE.

c. The negative binomial CI is more appropriate. The Poisson CI is overly optimistic, because it does not take into account the overdispersion.

CHAPTER 8

8.1 1. $z = 2.56$ (or McNemar chi-squared $= 6.53$), two-sided P-value $= 0.011$; strong evidence that low-birth rate cases are more likely than controls to be smokers.

8.3 For the subject-specific model, $\hat{\beta}$ is the log odds ratio conditional on the subject, whereas for the marginal model it is the log odds ratio for the combined sample, totaled over the subjects. The marginal model has $\hat{\beta} = \log[(359/785)/(334/810)] = 0.104$. The subject-specific model has $\hat{\beta} = \log(132/107) = 0.21$.

8.5 a. This is probability, under H_0, of an observed or more extreme result, with more extreme defined in the direction specified by H_a. In R, `binom.test(132, 239, 0.50, alternative = "greater")` gives P-value $= 0.06019$.

b. Mid P-value includes only half the observed probability, added to the probability of more extreme results.

8.7 a. Deviance $= 365.81$ with $df = 6$; more moves from (2) to (1) (std. resid. $= 6.0$), (1) to (4) (std. resid. $= 10.8$), (2) to (4) (std. resid. $= 12.7$) than we would expect if symmetry truly held.

b. Deviance $= 5.53$, $df = 3$; difference between deviances $= 365.81 - 5.53 = 360.3$, with $df = 6 - 3 = 3$, extremely strong evidence against marginal homogeneity (P-value < 0.0001 for testing H_0: marginal homogeneity). The small P-value mainly reflects the large sample size and is due to a small decrease in the proportion classified Catholic and increase in the proportion classified None or Other, with little evidence of change for other categories.

8.9 The difference of deviances, $402.2 - 2.1 = 400.1$ with $df = 6 - 5 = 1$, gives extremely strong evidence against marginal homogeneity (P-value < 0.0001). Subjects have a tendency to always respond more in the wrong direction for a married person having sexual relations with someone other than the marriage partner. For example, since $\log(\hat{\pi}_{14}/\hat{\pi}_{41}) = \hat{\beta}(4-1) = -8.6$, we conclude that π_{41} is higher than π_{14}; that is, the more favorable response is much more likely for premarital sex than for extramarital sex.

8.11 Deviance $= 13.8$ with $df = 11$, versus 346.4 with $df = 16$ for ordinary independence. Given a change in brands, the new choice of coffee brand is plausibly independent of the original choice.

8.13 kappa $= 0.389$; symmetry model fits well.

8.15 a. Residual deviance $= 2.08$ with df $= 6$. The estimates $(0.23, 0.28, -0.19, 1.09, 0)$ suggest that S. Williams ranks highest and Pliskova the lowest.

b. Estimated probability $= 0.75$ that S. Williams wins.

8.17 True.

CHAPTER 9

9.1 a. Sample proportion yes $= 0.86$ for A, 0.66 for C, and 0.42 for M. Marginal model: $\text{logit}[P(Y_t = 1)] = \alpha + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3$, where $t = 1, 2, 3$ refers to A, C, M, and $z_1 = 1$ if $t = 1$, $z_2 = 1$ if $t = 2$, $z_3 = 1$ if $t = 3$ (0 otherwise). For example, $e^{\beta_1 - \beta_3}$ is odds ratio comparing alcohol and marijuana use. Marginal homogeneity is $\beta_1 = \beta_2 = \beta_3$.

b. $\text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3$.

9.3 a. Marijuana: For $s_1 = s_2 = 0$, linear predictor takes greatest value when $r = 1$ and $g = 0$ (white males). For alcohol, $s_1 = 1, s_2 = 0$, linear predictor takes greatest value when $r = 1$ and $g = 1$ (white females).

b. Estimated odds for white subjects $\exp(0.38) = 1.46$ times estimated odds for black subjects.

c. For alcohol, estimated odds ratio $= \exp(-.20 + 0.37) = 1.19$; for cigarettes, $\exp(-0.20 + 0.22) = 1.02$; for marijuana, $\exp(-0.20) = 0.82$.

d. Estimated odds ratio $= \exp(1.93 + 0.37) = 9.97$.

e. Estimated odds ratio $= \exp(1.93) = 6.89$.

9.5 The GEE estimates show a linear time effect of 0.318 for the standard drug and $0.318 + 0.708$ for the new one. The difference in slopes of 0.708 has a GEE empirical

$SE = 0.136$, so the Wald statistic equals $(0.708/0.136)^2 = 5.2$ ($df = 1$, $P < 0.0001$) for comparing the slopes. Similar substantive results with scores (0, 1, 2).

9.7 a. Subjects can select any number of sources, so a given subject could have anywhere from 0 to 5 observations in table. Multinomial distribution does not apply to these 40 cells.

b. Estimated correlation is weak, so results not much different from treating 5 responses by a subject as if from 5 independent subjects. For source A, the estimated size effect is 1.08 and highly significant (Wald statistic = 6.46, $df = 1$, $P < 0.0001$). For sources C, D, and E, the size effect estimates are all weak and negative, roughly -0.2.

9.9 For a GEE approach with an independence working correlation structure for a cumulative logit model with constraint $\beta_E = 0$ for the environment, the estimates are $\hat{\beta}_C = -2.338$ ($SE = 0.121$), $\hat{\beta}_L = -0.465$ ($SE = 0.119$), $\hat{\beta}_H = -0.076$ ($SE = 0.116$). It appears that there is less support for spending on cities than for the other types of spending.

9.11 a. The estimated treatment log odds ratio changes from 0.00 to 1.41 as the initial response score goes from 10 to 75.

b. With no interaction, estimated treatment log odds ratio = 0.911 ($SE = 0.249$). Allowing interaction, the active treatment seems relatively more successful at the two highest initial-response levels.

9.13 They are independent, *conditional* on Y_{t-1}, but they are not independent marginally.

9.15 True.

CHAPTER 10

10.1 As q increases, $(\hat{\beta}, SE, \hat{\sigma}, SE)$ converges to (4.135, 0.713, 10.199, 1.792). For a given subject, estimated odds of belief in heaven are $\exp(4.135) = 62.5$ times estimated odds of belief in hell.

10.3 a. 0.4, 0.8, 0.2, 0.6, 0.6, 1.0, 0.8, 0.4, 0.6, 0.2.

b. $\text{logit}(\pi_i) = u_i + \alpha$. ML estimates $\hat{\alpha} = 0.259$ and $\hat{\sigma} = 0.557$. For average coin, estimated probability of head $= 0.56$. Predicted values are 0.52, 0.63, 0.46, 0.57, 0.57, 0.68, 0.63, 0.52, 0.57, 0.46.

c. Average distances are 0.22 in (a) and 0.08 in (b).

10.5 For $\hat{\beta}_A = 0$, $\hat{\beta}_B = 1.99$ ($SE = 0.35$), $\hat{\beta}_C = 2.51$ ($SE = 0.37$), with $\hat{\sigma} = 0$.

10.7 For the first model, for a given department, the estimated odds of admission for a female are $\exp(0.163) = 1.18$ times the estimated odds of admission for a male. For the second model, the estimated mean log odds ratio between gender and admissions, given department, is 0.176, corresponding to an odds ratio of 1.19. Because of the extra variance component, permitting heterogeneity among departments, the estimate of β is not as precise.

10.9 Effects in marginal models are smaller in absolute value than effects in GLMMs, with a greater difference when $\hat{\sigma}$ is larger. Here, the effect for GLMM is the same for each age group, but the effect diminishes more for the older age group in the marginal model because the older age group has a much larger $\hat{\sigma}$ in GLMM.

10.11 For a given subject, the estimated odds of response in category $\leq j$ on extramarital sex are $\exp(4.134) = 62.4$ times the estimated odds of response in those categories for premarital sex. For the marginal model, $e^{2.51} = 12.3$, so the estimated odds of response in category $\leq j$ on extramarital sex for a randomly selected subject are 12.3 times the estimated odds of response in those categories for premarital sex for another randomly selected subject. The estimate of β is much larger for the GLMM, since it is a subject-specific estimate and the variance component is large.

CHAPTER 11

11.1 Linear discriminant function for predicting I.virginica is -1.638(sepal) $+$ 3.152(petal). The classification table has 3 errors of each type.

11.3 b. As explained in Section 11.2.4, the choice $\lambda = 0.02$ may be overfitting. The tree mentioned in the text with $\lambda = 0.07$ seems more sensible and accords better with logistic modeling.

11.5 a. Q1: Is the subject's age >70? (3157 yes, 1497 no).
Q2: Is the subject's age >83? (931 yes, 2226 no).
Q3: Does the subject have dementia? (65 yes, 2161 no).
Q4: Does the subject have Parkinson's disease? (37 yes, 2124 no).

b. Misclassification cost is 819 if we predict that these subjects disenroll and it is $13(112) = 1456$ if we predict that they stay. Cost is lower if we predict that they all disenroll.

c. Tree correctly predicts $130/(130+195) = 0.40$ of those who actually disenrolled and $3426/(903+3426) = 0.79$ of those who remained.

11.7 One cluster (including DC) tends to vote Democratic and the other cluster tends to vote Republican (IN, NC, AZ, WY, UT, TX, SD, SC, OK, ND, NE, MS, KS, ID, AL, AK, GA, TN, MT, MO, LA, AR, KY).

11.9 The plot should look much like Figure 3.3 with width as the predictor. The GAM curve shows an increasing trend.

11.11 Compared to the null model, deviance decreases by $50.7 - 28.0 = 22.7$ on $df = 13$, for a P-value of 0.045. With $\lambda = 0.222$ from one analysis, the null model is selected.

BIBLIOGRAPHY

Agresti, A. (2013). *Categorical Data Analysis*, 3e. New York: Wiley.

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. New York: Wiley.

Allison, P. (2012). *Logistic Regression Using SAS: Theory and Application*, 2e. Cary, NC: SAS Institute.

Bartholomew, D.J., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*, 3e. Hoboken, NJ: Wiley.

Bartholomew, D.J., Steele, F., Moustaki, I., and Galbraith, J.I. (2008). *Analysis of Multivariate Social Science Data*. Boca Raton, FL: CRC Press.

Bilder, C.R., and Loughin, T.M. (2015). *Analysis of Categorical Data with R.* Boca Raton, FL: CRC Press.

Bishop, Y.V.V., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.

Burnham, K.P., and Anderson, D.R. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2e. New York: Springer.

Collins, L.M., and Lanza, S.T. (2009). *Latent Class and Latent Transition Analysis*. Hoboken, NJ: Wiley.

Fagerland, M.W., Lydersen, S. and Laake, P. (2017). *Statistical Analysis of Contingency Tables*. Boca Raton, FL: CRC Press.

Fitzmaurice, G., Laird, N., and Ware, J. (2011). *Applied Longitudinal Analysis*, 2e. Hoboken, NJ: Wiley.

Friendly, M., and Meyer, D. (2016). *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*. Boca Raton, FL: CRC Press.

Gelman, A., and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Goodman, L.A., and Kruskal, W.H. (1979). *Measures of Association for Cross Classifications*. New York: Springer-Verlag.

Gueorguieva, R. (2018). *Statistical Methods in Psychiatry and Related Fields: Longitudinal, Clustered, and Other Repeated Measures Data*. Boca Raton, FL: CRC Press.

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2e. New York: Springer.

Hedeker, D., and Gibbons, R.D. (2006). *Longitudinal Data Analysis*. Wiley.

Hoff, P.D. (2009). *A First Course in Bayesian Statistical Methods*. New York: Springer.

Hosmer, D.W., Lemeshow, S., and Sturdivant, R.X. (2013). *Applied Logistic Regression*, 3e. New York: Wiley.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning*. New York: Springer.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R*. New York: Birkhäuser.

Lloyd, C. J. (1999). *Statistical Analysis of Categorical Data*. New York: Wiley.

Magidson, J., and Vermunt, J.K. (2004). Latent class models. In: *The Sage Handbook of Quantitative Methodology for the Social Sciences* (ed. D. Kaplan), Chapter 10, 175–198. Thousand Oaks: Sage Publications.

Molenberghs, G., and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer.

Molenberghs, G., Fitzmaurice, G., Kenward, M.G., Tsiatis, A., and G. Verbeke, G. (eds.) (2015). *Handbook of Missing Data Methodology*. Boca Raton, FL: CRC Press.

Raudenbush, S., and Bryk, A. (2002). *Hierarchical Linear Models*, 2e Sage.

Simonoff, J.S. (2003). *Analyzing Categorical Data*. New York: Springer-Verlag.

Snijders, T.A.B., and Bosker, R.J. (2011). *Multilevel Analysis*, 2e. London: Sage.

Stokes, M.E., Davis, C.S. and Koch, G.G. (2012). *Categorical Data Analysis Using the SAS System*, 3e. Cary, NC: SAS Institute Inc.

Train, K.E. (2009). *Discrete Choice Methods with Simulation*, 2e. Cambridge: Cambridge University Press.

Tutz, G. (2011). *Regression for Categorical Data*. Cambridge: Cambridge University Press.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, New York: Wiley.

Wood, S.N. (2017). *Generalized Additive Models: An Introduction with R*, 2e. Boca Raton, FL: CRC Press.

Yee, T.W. (2015). *Vector Generalized Linear and Additive Models*, New York: Springer.

EXAMPLES INDEX

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e

SUBJECT INDEX

An Introduction to Categorical Data Analysis, Third Edition. Alan Agresti.

Companion Website: www.wiley.com/go/Agresti/CDA_3e